2010

Factors Influencing the Evaluation of Expert Witness Testimony in the Behavioral Sciences Under *Daubert*

Kenneth V. Heard
University of Rhode Island

FACTORS INFLUENCING THE EVALUATION OF EXPERT WITNESS TESTIMONY IN THE BEHAVIORAL SCIENCES UNDER *DAUBERT*

BY

KENNETH V. HEARD

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2010

ABSTRACT

Surveys on the evaluation of the scientific merit of expert witness testimony were mailed to 600 doctoral-level members of the American Psychological Association with professional experience in psychology and the law. Participants were asked questions related to their training, education, and professional experience, and questions aimed at clarifying participants' understanding of the error rate and general acceptance of methods. Participants were also asked to estimate the weight they would place on variables of potential relevance to the admissibility of expert witness testimony, and to estimate the weight they believe a judge would place on a subset of those variables. 126 surveys were returned in analyzable form. Results indicate that forensic psychologists are very consistent in their self-reported weights, and their estimated weights for judges. Participants report that they place moderate to great weight on virtually all variables and that they place greater weight on a wider range of variables than judges do, with the exception of face validity. The results are discussed in the context of legal, scientific, and philosophical views of scientific merit, and research on judgment and decision-making.

ACKNOWLEDGEMENTS

I owe so many people for their support, energy, care, and assistance on a variety of levels. I cannot adequately express my gratitude to any, nor to list them all here. My sincere gratitude goes to:

My family:

Sandi, you are my love and my harbor. Happy Anniversary!

Jacob, Miranda, and Theodore. "You are millions of locusts, and I am a blade of grass." I love you all for this privilege, and for the wonderful, talented, caring people that I watch you grow into every day.

Mom, Dad, and Toni. I would not have made it this far without you in so many ways.

The Brothers Heard: Nate, Matt and Tim.

Cap, Ruth and Sarkis. I am sad you are not here to share this with us all.

My committee members, present and past, who have extended me every possible consideration:

David Faust, Ph.D.

W. Grant Willis, Ph.D.

Donna Schwartz-Barcott, Ph.D., R.N.

Ellen Flannery-Schroeder, Ph.D.

Allan Berman, Ph.D.

Andre Ariew, Ph.D.

Larry Grebstein, Ph.D.

Sue Adams, Ph.D.

The following individuals, among others, were generous with their time and expertise during the development and conduct of this study:

Chief Justice Frank J. Williams, Rhode Island Supreme Court

Presiding Justice Joseph F. Rodgers, Jr., Rhode Island Superior Court

Presiding Justice Alice B. Gibney, Rhode Island Superior Court

Associate Justice Melanie W. Thunberg, Rhode Island Superior Court

Chief Judge Robert F. Arrigan, Rhode Island Workers' Compensation Court

Associate Judge Edward P. Sowa, Jr., Rhode Island Workers' Compensation Court

Chief United States Magistrate Judge Paul W. Grimm, U.S. District Court for the District of Maryland

Rhode Island Attorney General Jeffrey B. Pine

I must express my particular gratitude to my advisor, David Faust. I could not have asked for more in terms of support, mentoring, advocacy, and friendship. Your role in my intellectual and professional development cannot be over-stated.

The Faculty of the Department of Psychology for their support and patience.

Associate Dean Keith Killingbeck, for your support and advocacy.

Friends seen and unseen:

Carlos, Jack, Joe, Jon, & Marco. You help keep me sane. Or not, as the case may be.

Chip, Jim, and John. I miss you all.

Jean Maher for years of support and friendship.

Dr. Joan Chrisler, for her steadfast mentorship, support, and friendship.

Last, I must express my additional gratitude to my 6' 8" one-ton brother, Dr. Nathan J. Heard, for decades of loving intellectual Fight Club.

TABLE OF CONTENTS

LIST OF TABLES

Title of the Study

Factors Influencing the Evaluation of

Expert Witness Testimony in the Behavioral Sciences

Under *Daubert*

STATEMENT OF THE PROBLEM

The outcome of legal proceedings in Federal, and in many state, jurisdictions may be determined in whole or part by the admissibility of expert witness testimony under the standard established by the United States Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals* in 1993. Under that ruling, the admissibility of proposed expert witness testimony is determined by the presiding judge in a given case who, acting as a gatekeeper, has an affirmative responsibility to exclude testimony that is irrelevant, unreliable, prejudicial, or that would otherwise not be deemed helpful to the trier of fact (i.e., the judge or jury). Consequently there has been increased attention among judges, attorneys, and individuals who would serve as expert witnesses to issues related to the evaluation of scientific evidence.

The appraisal of theories and methods for the purpose of determining their scientific standing can be problematic, particularly within the "softer" sciences. This is certainly true for judges and attorneys who may lack training in science and the scientific method, but may also be true for professionals within a given field of expertise when faced with a diversity of studies of varying methodological quality and findings from different studies that are incompatible or contradictory. Research in

human judgment suggests that individuals making such appraisals will employ a small number of judgmental heuristics or cues. Given the different training, education, experience, and roles that attorneys, judges, and experts possess, it is unlikely that they will ascribe the same significance or weight to different cues they use in their evaluations of scientific trustworthiness or merit.[1] The degree to which members of the same professional class agree on the cues that tend to best reflect the scientific status of proposed expert witness testimony may also vary dramatically.

JUSTIFICATION FOR AND SIGNIFICANCE OF THE STUDY

The Expert Witness

The role of the expert witness in Federal courts is defined by the Federal Rules of Evidence Rule 702, which states

> If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

Thus, an expert witness is one who, by virtue of education, training, and experience, is able to assist a judge or jury in understanding scientific, technical, or other specialized

[1] When referring to scientific merit or trustworthiness, *Daubert* refers to the "reliability" of evidence. Reliability in the legal sense differs from "reliability" as commonly understood in social science. In the legal system, "reliable" means having a valid factual foundation rather than referring to the precision and reproducibility of scores or results. Thus, the court's use of reliable is closer to the psychologist's use of "valid." To avoid confusion, the terms "trustworthiness" or "merit" will be generally substituted for the terms "reliability" and "validity" as understood by attorneys and psychologists, respectively.

knowledge that is beyond the knowledge and experience of the average person (Heard & Faust, 2002).

Procedurally, expert witnesses differ from lay (or fact) witnesses in that they may offer opinions or interpret the facts in evidence, provided that any opinions offered can be demonstrated to be supported by facts in evidence and based on scientifically validated methods that have been appropriately applied to the case at hand. Expert witnesses (unlike lay witnesses) may also testify to facts that would otherwise be considered inadmissible as hearsay,[2] provided that the material is of the sort normally used to form opinions within the expert's field (as per The Federal Rules of Evidence Rule 803). To illustrate, a lay witness might testify, based on his or her direct observations, that a plaintiff was tremulous and crying in the emergency room following a car accident, but would not normally be allowed to discuss anything not witnessed personally, draw inferences, or offer any conclusions. In contrast, a clinical psychologist acting as an expert witness might testify at length about the etiology, course, treatment, and functional consequences of Posttraumatic Stress Disorder, the relevant scholarly literature, and discuss details from various second-hand sources of information, such as treatment records, accident reports, and interviews with the plaintiff's family members. They likely would also be allowed to discuss inferred inner mental states, such as anticipatory anxiety. Psychologists testify as expert witnesses in a variety of contexts, including those in which mental state, etiology,

[2]Although there may be different procedures and standards employed by different courts, "hearsay," meaning evidence based on the reports of others rather than the personal knowledge of the witness, is generally not admissible as testimony during trial and therefore cannot be heard or considered by the trier of fact.

diagnosis, or consequences of psychological disorder or injury is at issue in criminal and civil venues (Heard & Faust, 2002; Melton, Petrila, Poythress, & Slobogin, 1997).

Prior to testifying, a potential expert witness must be proffered to the court and certified as an expert. The qualifying of an expert by the court is conducted through examination of the proffered witness, under oath, by attorneys from one or more parties to the litigation in a process known as *Voir Dire*. *Voir Dire* is a 'trial within the trial' used to determine the competence, qualifications, and knowledge of the expert, and to assess the relevance of the expert's testimony. This process usually involves a presentation and review of the witness's credentials and of the subject matter to be addressed during that expert's testimony, and serves to define the scope of the witness's expertise for the purposes of the trial at hand (Heard & Faust, 2003). In addition, the expert's ability to communicate his or her knowledge to a lay audience may be considered by the judge when deciding the admissibility of the testimony. An expert who has impeccable credentials and is fully competent in her or his area, but who cannot speak so as to be understood (e.g., is disorganized, mumbles, or unable to translate abstract or technical terminology into plainer speech) fails the test of "helpfulness," in that they cannot assist the trier of fact understand the evidence in question (Faust, Grimm, Ahern, & Sokolik, In Press).

As an expert witness's testimony may be important or even critical in a legal dispute, an opposing party may challenge the witness's expertise, credentials, or credibility during *Voir Dire* in an effort to have her or his testimony limited or even excluded entirely. For example, a psychologist presented to testify on the diagnosis and prognosis of Post Concussion Syndrome might be required to demonstrate

specific training and expertise in neuropsychology beyond that typical of the average clinician. Even if admitted as an expert, the scope of the witness's testimony may be limited if the judge's opinion is that certain areas lie beyond that individual's expertise or are not relevant to the matter at hand. Suppose a plaintiff has incurred a brain injury with lesions identified through radiological imaging. A particular judge could determine that a particular neuropsychologist is not qualified to offer interpretation of neuroimaging, but allow testimony about the effects of those lesions, provided their existence has been established by other records or other testimony by an individual deemed qualified. Likewise, that same psychologist might be precluded from testifying on subjects related to Posttraumatic Stress Disorder in the same case, despite acknowledged qualifications, if PTSD was not at issue.[3]

The certification of an expert prior to her or his testimony at trial does not necessarily reflect the strength or trustworthiness of the testimony she or he will be called on to present. A judge may identify potential problems with the witness's credentials or credibility, or with elements of the proposed testimony, but determine that the testimony meets a minimum threshold for admissibility. In such a case, issues raised at *Voir Dire* may be raised again during cross-examination of the expert witness at trial. The attorney for the opposing party may also raise a variety of additional issues during cross examination not previously addressed during the certification process. Weaknesses or problems so exposed 'go to the weight' that the trier of fact may ascribe to the testimony. That is, a judge may rule that an expert witness is

[3] It is also possible that an individual with expert credentials may be called to testify as a lay witness, speaking only to what he or she has personally done or observed. Under these circumstances, he or she would generally fall under the rules that apply to lay witnesses despite the relevance of any expertise he or she may possess.

qualified to testify and that the testimony in question is potentially relevant and helpful, but the jury could find that the expert's testimony is not persuasive.

The *Frye* Standard

In modern times, within Federal Court (and jurisdictions incorporating Federal standards) *Frye v. United States* (1923) has been the most significant pre-*Daubert* ruling on the matter of expert testimony. In *Frye*, the defendant in a criminal case sought to have an expert provide testimony that the defendant's assertion of innocence was truthful based on the experimental use of blood pressure measurements. The court ruled the expert's testimony inadmissible because it was not based on techniques that were generally accepted within the relevant scientific community.

Under *Frye*, the judge in essence does not directly appraise the scientific basis of testimony, but makes a decision as to the degree to which that testimony is accepted within the applicable field. *Frye* is thus excessively conservative at the frontiers of science where novel discoveries of real merit have not gained broad acceptance, yet excessively liberal in cases where an expert community (for whatever reason) accepts theories or techniques that are of questionable merit or even demonstrably false (c.f., Gorman, 1999). To complicate matters further, even when there is a clear majority opinion within a field of expertise, a judge may defer to the opinion of a substantial or "respected" minority within that field and admit or reject testimony despite the majority view. Additionally, the boundaries of the relevant scientific community may be drawn differently or disputed depending on the specifics of the matter at hand, and what constitutes the relevant community of professionals or a "substantial" minority is also open to debate.

For example, the use of the Rorschach Inkblot Test[4] in psychological assessment, and as a foundation for courtroom testimony, has been subject to considerable debate in the professional literature (c.f., Garb, 1999; Grove, Barden, Garb, & Lilienfeld, 2002; Lilienfeld, Wood, & Garb, 2000; McCann, 1998; Ritzler, Erard, & Pettigrew, 2002; Rosenthal, Hiller, Bornstein, Berry, & Brunell-Neuleib, 2001). Concern among Rorschach proponents for the method's viability in forensic contexts is highlighted by a recent position paper by the *Society for Personality Assessment* which asserts that the Rorschach "meets the variety of legal tests for admissibility, including validity, publication in peer reviewed (sic) journals, and acceptance within the relevant professional community" (SPA Board of Trustees, 2005, p. 221). Without advocating for a position on the scientific status of the Rorschach, one might reasonably ask whether the "relevant professional community" in this case is the body of trained Rorschachers, the membership of the *Society for Personality Assessment*, those professionals who engage in the practice of psychological assessment, or a larger body of individuals with professional expertise in psychometrics, testing, and assessment. Further, depending on where that boundary is drawn, at what point might one reasonably define the body of Rorschach advocates as a "sizeable minority" of the larger field? Such circumstances, difficult to resolve within psychology, may present an intractable problem for the judge ruling on the admissibility of testimony.

[4] This procedure involves the presentation of a series of inkblots to an individual, who then describes what he or she perceives in the card, much in the way individuals may see different things when gazing at clouds. In theory, as the inkblots are ambiguous and the range of possible responses infinite, the individual's responses should reflect aspects of his or her personality and mental processes.

In practice, *Frye* opens the courts to a wide variety of professed experts offering testimony that often lacks an adequate scientific foundation and, in some cases, testimony properly described as pseudo- or junk science. Concern about the proliferation of junk science in the courtroom has resulted in mounting pressure within the legal system and Congress for reform (Dixon & Gill, 2001; Schuman & Sales, 1999).

In comparison to *Frye*, under which the judge defers to the opinion of the scientific community, the Federal Rules of Evidence Rule 702 charges the judge with the task of determining whether or not expert testimony of a scientific, technical, or specialized nature has a sound scientific basis. The role of the judge is explicitly that of gatekeeper under the Federal Rules of Evidence. Rule 702 and *Frye* are therefore sometimes at odds within the legal system as each requires the judge to take a different role with respect to the admissibility of expert witness testimony. This circumstance obtained within Federal jurisdictions, and those state jurisdictions following the Federal Rules, until the U.S. Supreme Court's ruling in *Daubert. v. Merrell Dow Pharmaceuticals* (1993).

The *Daubert* Standard

In *Daubert*, the defendant in a civil case sought the exclusion of testimony by a plaintiff's expert alleging a causal relationship between certain birth defects and the anti-nausea drug Bendectin, which was prescribed for sickness during pregnancy. The Court ruled in favor of the defendant, finding that the testimony in question lacked a sufficient scientific foundation to be admitted, citing factors such as lack of supportive studies published in peer-reviewed journals. In doing so, the Court specified that the

Federal Rules of Evidence (and the gate-keeping function of the presiding judge in a given case) supersedes *Frye* (although judges may incorporate general acceptance, or the so-called "*Frye* standard," in making their decisions).

Subsequently, in *Kumho Tire v. Carmichael* (1999), the U.S. Supreme Court affirmed that the trial judge's gatekeeping function applied to all forms of expert testimony, not only those explicitly identified as "scientific" or as using the scientific method. In *Kumho*, an expert on automotive tire tread wear was excluded despite the argument that the testimony was based on specialized technical knowledge and his years of experience in his field rather than the scientific method or formal experimentation, and, thus, not subject to the *Daubert* standard. *Kumho* acknowledges explicitly that the *Daubert* criteria may not apply equally well to all fields of expertise, and that experience alone may provide a sufficient basis for reliable exert testimony in some circumstances. However the Rules, post-*Kumho,* specify that "if the witness is relying solely or primarily on experience, then the witness must explain how that experience leads to the conclusion reached, why that experience is a sufficient basis for the opinion, and how that experience is reliably applied to the facts" (Committee Notes on Rules - 2000 Amendment). *Kumho* had wide-sweeping implications because various experts, who had claimed scientific status when *Frye* was in effect, shifted their emphasis to expertise acquired through applied practice or experience when *Daubert* came into force.

The other major modification to *Daubert,* coming from *General Electric v. Joiner (1997*), addresses the proper basis for the appeal of cases to a higher court. Once a decision has been rendered by the trier of fact in a given case, a party to the

litigation who is dissatisfied with the outcome may seek to have the decision reversed or retried by a higher court. In doing so, the appellant seeks to prove that the outcome of the original trial was unjust, because the judge committed some action that does not adhere to legal guidelines, and that, but for this error, the outcome could have been different. Some examples of reversible error on the part of the presiding judge may include procedural violations, manifest evidence of bias against some party to litigation, or the issuance of improper instructions to the jury. The exclusion or admission of an expert for inappropriate reasons constitutes another class of potential reversible error and this issue is often raised as the basis for appeal of decisions when expert witness testimony has been proffered.

Although the law provides rules and principles governing judicial procedure, it is recognized that all possible circumstances cannot be anticipated and codified in advance. Judges are required to exercise a degree of discretion in interpreting and applying the law to the particular facts of individual cases. Even if there are no gross violations of legal dictates, a party to litigation may believe that a judge's decision is unreasonable (that is, reasonable persons with an understanding of the law and the facts of the matter could disagree with the ruling) and may appeal on the grounds that the judge has committed an "abuse of discretion." If the appellant court finds that the lower court's decision meets the test of reasonableness, or that the alleged abuse of discretion was harmless (i.e., did not alter the outcome of the trial), then it will affirm the previous decision. If, however, the appellate court finds that reasonable persons reviewing the specific facts in light of the relevant law could reach a different conclusion, then it may rule that an abuse of discretion has occurred. In the latter case,

the appellant court may overturn the original decision and substitute its own, or may remand the matter back to the original court for further litigation.

In *Joiner*, the plaintiff sought admission of testimony that the plaintiff's occupational exposure to PCBs caused his small-cell lung cancer. The defense challenged the admissibility of the expert on the grounds that the proposed testimony lacked a valid scientific foundation. The District Court agreed and excluded the expert, stating that the expert's opinions did not rise above "subjective belief and unsupported speculation." The plaintiff appealed the case, arguing that the judge had wrongfully excluded the expert's testimony and, in doing so, had undermined his case. The matter was eventually argued before the U.S. Supreme Court. At issue was the question of the standard for the appellate court's review of the original decision to exclude the expert. The plaintiff argued that the appellate court should complete a novel review of the proffered testimony and make a determination as to its scientific merit. The defendant argued that appellate review should be limited to the narrow question of whether or not the judge, in deciding to exclude the testimony, had abused his discretion. The Supreme Court found in favor of the defendant, ruling that appeals of *Daubert* decisions excluding or admitting testimony should be heard as issues of abuse of discretion as opposed to a more detailed appellate review of the scientific merits of the testimony in question. Although there have been various other decisions within the courts which have discussed issues related to *Daubert* and its application, including one argued before the U.S. Supreme Court (*Weisgram v. Marley Co.*, 2000), these three cases (*Daubert, Kumho*, and *Joiner*) are commonly know as the "*Daubert* Trilogy."

Daubert and its derivations now apply in Federal courts and fully or partially in states that have adopted these (or similar) standards. *Daubert* challenges to expert admissibility have been made in a variety of areas including medicine, engineering, forensic science, and social science. Such efforts to exclude or limit an expert's testimony are often based on the adequacy of that person's training, experience, and credentials, but experts may also be challenged on the scientific foundation of their methods, the application of their methods in the specific case at hand, or the relevance of their testimony to the facts at issue.

The Evaluation of Testimony Under *Daubert*

There has been great variance in the application of *Daubert* due partly to the wide latitude in weight and interpretation provided for in the original ruling. In explaining its decision in *Daubert*, the Court provided a non-exhaustive, non-ordinal, and non-hierarchical list of issues to be considered when evaluating the validity of expert testimony. The four criteria discussed were:

1) General acceptance within the relevant scientific community (as per *Frye*).
2) The known or potential error rate[5] of the theory or technique.
3) The degree to which the theory is testable and falsifiable and has been subjected to formal study.
4) Publication in peer-reviewed journals or outlets.

[5] Although the text of the *Daubert* decision does not explicitly specify that "error rate" refers to the accuracy of methods, i.e., the number of correct versus incorrect decisions, this is how attorneys, judges, and forensic psychologists generally understand the term in that context. The term could be interpreted by some as relating to Type I and Type II error from statistical significance testing in research, but this is not the sense in which the term is typically used in legal proceedings when the accuracy of assessment and predictive methods is at issue. There may be circumstances in which Type I and Type II error are at issue in legal proceedings, but the term is used to connote accuracy in this paper unless explicitly specified otherwise.

The relative weight of these criteria is not specified, nor is any guidance provided for the case in which only some of the criteria can be evaluated, or when they are inconsistent. For example, the Court did not specify how to rule if there is rigorous peer-reviewed literature supporting a theory, but it enjoys only limited acceptance. Furthermore, the list is specifically described as non-exhaustive. At their discretion, and based on the arguments presented to them, judges may apply any number of additional criteria instead of, in addition to, or in various combinations with those specified in the ruling. For example, under *Daubert*, the credibility of an expert may be an important, or even deciding, factor in the judge's decision on admissibility, independent of the scientific basis of the proffered testimony as evaluated by the four identified criteria. Finally, as discussed in *Kumho v. Carmichael*, not all *Daubert* factors may apply equally, or at all, to some areas of expertise.

Recently, Rule 702 has been modified to include the following *Daubert*-related standards for admissibility (O'Connor & Krauss, 2001):

1) Is the testimony based on sufficient facts or data?
2) Is the testimony the product of reliable [i.e., trustworthy] principles and methods?
3) Has the witness applied the principles and methods reliably [i.e., in a trustworthy way] to the facts of the case?

The current edition of the Rules, in providing further explanation and clarification of Rule 702 (Committee Notes on Rules - 2000 Amendment), explicitly states that other factors may be relevant in establishing the validity of testimony, and specifies as examples:

1) Whether experts are "proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying."
2) Whether the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion.
3) Whether the expert has adequately accounted for obvious alternative explanations.
4) Whether the expert "is being as careful as he would be in his regular professional work outside his paid litigation consulting."
5) Whether the field of expertise claimed by the expert is known to reach reliable [i.e., trustworthy] results for the type of opinion the expert would give.

The Impact of *Daubert* in the Behavioral Sciences

Daubert touched off a firestorm of debate in the legal and scientific communities, including substantial discussion of how the standards would be interpreted and implemented by the courts. On the one hand, *Daubert* in some ways liberalizes the standard of admissibility as an expert may be allowed to present evidence that is not necessarily endorsed by the broader scientific community. This can serve to widen the range of potential topics for consideration by the Court, and recognizes that science is constantly evolving. On the other hand, by requiring judges to critically evaluate the scientific status of proposed testimony, *Daubert* has the potential to function as a barrier to junk science entering into the legal process. The

gatekeeping function of judges can require that experts defend their methodology and its application in ways that *Frye* did not necessitate. Not only should testimony be based on a demonstrable foundation of scientific research, but an expert may also be challenged for the inaccurate, improper, or untested use of otherwise validated methods. Thus, conclusions drawn from a well-validated neuropsychological test battery might be judged invalid, and thus excluded, if it can be demonstrated that the administration was rife with scoring errors or deviations from standard procedures. Likewise, a test which is well-validated for some purposes in some populations may lack any validation studies for the specific application or population under consideration.

Within psychology, the response to *Daubert* has been largely from the armchair, involving articles predicting the possible effects of the ruling (Grove & Barden, 1999; Lipton, 1999; Penrod, Fulero, & Cutler, 1995; Shuman & Sales, 1999; Youngstrom & Busch, 2000; Zonana, 1994), proscriptions and recommendations for the psychologist working within the framework provided by *Daubert* (Brodsky, 1999; Ceci & Hembrooke, 1998; Goodman-Delahunty, 1997; Gutheil & Stein, 2000; Rotgers & Barrett, 1996), and debate about the degree different techniques, diagnoses, and procedures meet (or fail to meet) the *Daubert* standard (Dyer & McCann, 2000; Heath, 2000; Lally, 2001; McCann, 1998; Pope, 1998; Reed, 1996; Rogers, Salekin, & Sewell, 1999; Saxe & Ben-Shakkar, 1999; Vallabhajosula & van Gorp, 2001). However, despite the lack of empirical study, there is a growing industry of individuals and firms which advertise themselves explicitly as *Daubert* experts.

The admissibility of psychological and social sciences testimony has been decided under *Daubert* in a growing number of published decisions, which include both civil and criminal matters (Bursoff, 1999; see also Faust, Grimm, Ahern, & Sokolik, In Press). Examples of the types of evidence proffered to date have included (broadly) testimony on syndromes, behavior profiles and diagnostic techniques related to sexual abuse/victimization; eyewitness identification; theory and experimental data related to civil rights and discrimination; the prediction of dangerousness and causal or mitigating factors in violent behavior; issues of mental state or condition; and "truth-telling" techniques (i.e., polygraph). The Federal Rules of Evidence note that "a review of the caselaw after *Daubert* shows that the rejection of expert testimony is the exception rather than the rule" (Committee Notes on Rules - 2000 Amendment). However, across a sample (1993 – 1999) of approximately 80 cases in which psychological testimony was subjected to a *Daubert* challenge, the rate of rejection exceeded 50% (Bursoff, 1999). A more recent tabulation of cases, covering 59 published and unpublished appellate decisions in Federal courts between 2000 and 2006 found a rejection rate of 64% (Nordberg, 2006).[6]

Given that not all court decisions are published, and that attorneys may withdraw potential witnesses they believe to be vulnerable prior to a final hearing before the presiding judge, the overall impact of *Daubert* on potential testimony by psychologists is certainly broader than these tabulations would suggest. Although some areas may be more vulnerable to a *Daubert* challenge than others, the possibility of exclusion is a legitimate concern to forensic psychologists, attorneys, and litigating

[6] This sample may be skewed toward rulings of exclusion due to the sole inclusion of appellate cases. To the extent that the qualifications and methods of experts are not controversial or in some way vulnerable to challenge, attorneys may be less likely to appeal (Nordberg, personal communication).

parties where the expert testimony of psychologists is significant. Exclusion of a key expert, especially late in proceedings and beyond the point at which a replacement is possible, may spell the end of the case. Claims (and the defenses against them) based on psychological testimony may be substantially (and often critically) undermined when psychological experts are excluded from testifying.

The large majority of civil and criminal cases are either settled or dropped prior to trial, and settlement during trial is not uncommon. Cases that end up in trial are often those in which there is at least some probative evidence supportive of each side. Where the plaintiff's (or prosecutor's) case is overwhelmingly strong, settlement is often the best course of action for the defense. When the opposite holds true, plaintiff's attorneys (or prosecutors) will judge the case as weak and likewise tend not to proceed. Consequently, although the strength of a case from a legal perspective may not be directly related to the scientific merit of potential expert testimony, a *Daubert* challenge may well occur in circumstance in which the evidence is mixed.

The attempt to exclude some or all of an expert witness's testimony often involves a considerable expenditure of effort and financial resources, is not without risk, and the cost of failure may be significant. In order to seek the exclusion or limitation of an expert's testimony, the attorney will most likely have to reveal various elements of her case formulation, particular facts she considers to be important, or potential elements of cross-examination strategy to the opposition. Should the attempt fail and the expert be allowed to testify, the expert and opposing attorneys will likely be forewarned and better prepared to modify and strengthen the expert's testimony. For example, if, in a failed motion to exclude an expert, an attorney reveals that a

particular test was scored incorrectly, or that there is research contradicting certain of the expert's conclusions, the opposing attorney and expert can then prepare for these specific lines of questioning and often mitigate any potential damage.

The human and financial costs of failure to accurately gauge the probability of success for a potential *Daubert* challenge can be great for both litigators and the involved parties. Worse, by definition, an expert witness possesses knowledge or expertise that the judge does not, and judges and attorneys may lack the scientific and methodological training that are required to properly evaluate the scientific literature. This is a particularly difficult problem within psychology, where the quality of methodological rigor and sophistication varies dramatically (Meehl, 1973; Meehl 1978/1991), scientific progress is often difficult at best (Meehl, 1978/1991; 1997), and there is often a substantial gap between research and practice (Gorman, 1999; Grove & Meehl, 1996). The summation and evaluation of research in any given area of psychology is, in itself, often problematic, even for experts within the field (Meehl, 1990), and lawyers may often approach behavioral science research and findings with a different perspective and language than the practitioner (Faust & Heard, 2003a, 2003b; Meehl, 1971/1991). Thus, although in principle it is a laudatory step for the Court to require adequate scientific support for an expert's opinions, a potentially troubling circumstance has been created.

Humans often must make decisions based on a series of fallible (probabilistic) indicators or pieces of data (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996). The task of the judge deciding the admissibility of expert psychological testimony is clearly of this nature, as it is of the attorney or expert witness sizing up the probability

that testimony will be admitted by a given judge. Various features of the credentials of a potential expert and the scientific merit of the proffered testimony are presented and argued, and the true state of these credentials and scientific status may be obscured or illuminated by the skill, preparation, and understanding of the attorneys involved in the case. In addition, judges must often make these challenging decisions in the absence of broader context that might assist their assessment of the underlying scientific merit (Grim, In Press). Momentarily putting aside deeper issues in the philosophy of science, clearly, none of the *Daubert* guidelines are perfectly predictive of the true scientific status of the ideas or concepts to be presented in testimony. Judges will at times admit testimony that should have been excluded based on a proper appraisal of scientific merit and vice versa. Consequently, the attorney attempting to predict the outcome of a *Daubert* challenge is in a difficult position: one who may lack expertise in scientific methodology (and thus may not be able to best formulate and articulate the true scientific status of testimony) must anticipate the ruling of a judge who also may lack expertise in scientific methodology. Furthermore, the expert witness, despite good faith and diligence, may also misjudge the strength of the scientific evidence underlying his theories, methods, and conclusions. As judges and attorneys must rely upon the expert to assist them in evaluating scientific status, they are in a true quandary in that they must make decisions based on probabilistic indicators that are in turn presented and explained by experts whose interpretation of these indicators is only probabilistically related to the true status of the science involved.

Research on *Daubert*

Although there has been copious discussion of *Daubert*, its merits, and implications, minimal scientific research has been conducted to date on the ruling, its application, and impact. Each of the available studies suffers from methodological limitations, including, in each case, issues relating to the representativeness of the samples. Johnson, Krafka, and Cecil (2000) and Krafka et al. (2002) provided an analysis of a 1998 survey of 303 Federal district court judges on their experiences with expert testimony in civil cases, their most recent civil trial in which admissibility was at issue, and opinions about admissibility post-*Daubert*. This study included analysis of admissibility rulings from the 297 Federal civil trials identified by the surveyed judges. As a follow-up, the authors also surveyed lead attorneys from these cases.

According to this research, medical and mental health experts comprised together the largest category of proffered testimony (43%). Of these, 18% (or nearly 8% of all expert witnesses) were clinical psychologists, psychiatrists, social workers or counselors. Not included in these numbers are a small proportion of other social and behavioral scientists aggregated within "other scientific specialties" in the report.

Judges reported that in almost half of the cases (46%), admissibility was not in dispute, and that they rarely (3% of cases) raised the question of admissibility themselves under these circumstances. When admissibility was disputed by some party, 59% of judges reported that they had allowed the testimony without limitation, a decline from 75% in a similar 1991 (pre-*Daubert*) survey. The proportion of cases in which psychological testimony (or testimony for allied fields) was disputed or

excluded was not reported. For those cases in which testimony was excluded, judges cited the following reasons in at least 15% of cases:

Proffered testimony was not relevant	47%
Proffered witness was not qualified	42%
Proffered testimony would not assist the trier of fact	40%
Facts or data underlying testimony not reliable [i.e., trustworthy]	22%
Testimony more prejudicial than probative[7]	21%
Principles and methods underlying testimony not reliable [i.e., trustworthy]	18%

The reasons for exclusion are provided for all cases in the aggregate, and it is not possible to determine whether these reasons apply to psychological testimony in the same proportions reported. Judges could also have cited more than one reason per case.

In addition, judges rated the frequency with which they encountered problems with expert witnesses in civil cases, some of which would seem clearly relevant to the issue of admissibility (e.g., expert testimony not comprehensible to the trier of fact) and others of which less evidently so (e.g., attorneys unable to adequately cross-examine experts). Of those seemingly relevant problems, judges rated "experts abandon objectivity and become advocates for the side that hired them" and "expert

[7] Evidence or testimony may be excluded if the probative value of the evidence is substantially outweighed by the danger that it may create unfair prejudice, cause confusion of the issues, or mislead the jury. For example, graphic crime-scene photographs of a particularly heinous crime might be considered potentially prejudicial against a defendant if simply viewing them would predispose a jury to convict out of a sense of disgust and outrage, without due consideration of the facts at issue. Likewise, should a jury be overly impressed by an expert's credentials, publication record, and use of technical jargon and thus not give adequate critical consideration to that expert's conclusions, it could be argued that that expert's testimony is more prejudicial than probative, particularly if the testimony lacks a sound scientific foundation.

testimony appears to be of questionable validity or reliability" as among the most frequent.

Gatowski et al. (2001) surveyed 400 state court judges to elicit their opinions about decision-making under *Daubert*. The overwhelming majority (91%) reported that the gatekeeping role was appropriate, and approximately half (52%) felt that their education had adequately prepared them for this role. However, 96% reported that they had not received specific instruction in the general methods and principles of science, and few were able to produce an adequate explanation of "falsifiability" or "error rate" (based on Gatowski et al.'s evaluation of unstructured responses) as specified in *Daubert*. On another item, almost all respondents evaluated the four guidelines that *Daubert* specifies for evaluating the trustworthiness or merit of proffered testimony as being useful. There was little consensus on how to weigh the *Daubert* criteria, although "general acceptance" was slightly favored over the others.

These findings potentially create a puzzling paradox, in that almost all judges described "falsifiability" as useful in evaluating testimony, yet, according to the authors, very few were able to provide an adequate definition of the concept. How can a criterion judges may not understand be helpful to them in rendering a decision? It may be the case that judges may have an understanding of falsifiability that is different than that of scientists or philosophers of science but that is still relevant to the more general question of scientific merit (e.g., has a method been subjected to adequate testing). However, it is not necessarily the case that understanding is a prerequisite for the development of a subjective sense that information is "useful." For example, if a judge employs a "vote-counting" method of evaluating testimony, then the witness's

credible assertion under cross-examination that the theory is "falsifiable" might be sufficient to produce a mental checkmark next to that criterion. Although the judge does not evaluate whether or not the theory is, in fact, falsifiable, and may not be able to define falsifiability, the assertion, especially if not contested effectively, may be sufficient to carry the day.

Such self-reports of decision-making strategies are open to question, however. Self-reports of how pieces of information, or "cues," are combined and weighed can be compared to how cues are combined and weighed when analyzed statistically. Research on insight into how cues influence decision-making often shows that there is a sizable gap between subjective impressions and objective measures of cue use (Dawes, 1979; Goldberg, 1970). Thus, for example, although judges may believe that they give nearly equal weight to all four *Daubert* guidelines, statistical analysis of decisions might demonstrate that some criteria receive minimal weight, others great weight, or even that status of a single factor (e.g., the presence or absence of peer-reviewed literature) is decisive. Individuals, regardless of their capabilities as decision-makers, may make decisions with less, or far less, than perfect insight into how those decisions are made.

Dahir, Richardson, Ginsburg, Gatowski, Dobbin, and Merlino (2005) subsequently published previously unreported results from the above-discussed Gatowski et al. (2001) study. Judges were asked an open-ended question about the factors they considered when determining the admissibility of evidence. Although there were no instructions to provide an ordinal list, the authors coded only the first four factors listed by judges, presuming that the response order would reflect the

importance judges ascribed to the factors they considered. Of the 216 judges who provided codable answers and who had experience with considering the admissibility of psychological syndromes, the following were the most frequent responses:

Qualifications of the expert	48%
Relevance and foundational issues	29%
General acceptance	25%
General appraisal of reliability [i.e., trustworthiness]	11%
Credibility and other expert characteristics	10%

Of the *Daubert* criteria for evaluation, "general acceptance" was frequently noted in the responses coded. Of the other *Daubert* criteria, 16 judges (7%) mentioned peer review and publication, two mentioned falsifiability, one mentioned error rate, and eight judges noted "all *Daubert* criteria." The authors conclude that *Daubert* has not had a significant impact on admissibility decisions, and that judges are more comfortable with the long-standing *Frye* standard.

The methodological concerns described for the Gatowski et al. study above apply equally to Dahir et al., limiting the conclusions that can be drawn from this secondary analysis. Further, one could reasonably question the authors' assumption that the order in which judges listed the criteria they considered necessarily reflected their subjective rankings or weights. The decision to code only the first four responses given by judges limits the interpretability of the frequency counts reported as one has no way to know which additional factors, or the number of additional factors, judges may have reported.

Groscup et al. (2002) collected 693 published state and federal appellate court decisions in which the admissibility of expert testimony was a substantive issue and concluded that satisfaction of the *Daubert* criteria for establishing scientific merit was not a good predictor of admissibility decisions. Indeed, although there was decreased discussion of *Frye* and "general acceptance" and an increase in the number of words devoted to discussion of *Daubert*, the other *Daubert* criteria were rarely mentioned specifically in the texts of the decisions they analyzed. They suggested that courts are engaging in more critical review of expert testimony in general based on increased word count, despite the lack of specific reference to the basis for admission or exclusion or testimony. However, they observed no overall difference in the rate of admissibility in the 5 ½ years pre- and post-*Daubert*. It seems unlikely that the level of critical analysis of testimony could have increased meaningfully without some corresponding change in admissibility rates. It is theoretically possible that testimony that would have been previously admitted under *Frye* began to be excluded post-*Daubert* due to inadequate scientific support, while testimony less well accepted, but with greater scientific merit, began to be admitted at a comparable rate. This seems implausible.

Groscup et al.'s results are limited due to significant methodological flaws. The selection of appellate case is perhaps understandable, as a greater proportion of appellate decisions are published and accessible via the type of search engines they utilized for identifying cases for analysis. However, criminal appellate cases in general are probably not representative of the broader spectrum of litigation subject to *Daubert*. Nearly 98% of the admissibility challenges were made by convicted

criminal defendants, and the overwhelming majority of experts challenged on appeal were government/prosecution witnesses who had been admitted at the time of the original trial. Problems derive from these characteristics of the sample. First, most of the experts in question were admitted by the lower court and testified for the party that prevailed under the "beyond a reasonable doubt" standard for conviction. Second, in criminal cases involving severe sentences (most notably the death penalty), appeals are de rigeur and frequently involve a "kitchen sink" approach in the hope of finding a reason for reversal of the conviction. That is, challenges to the admissibility of expert testimony are more likely to be made without regard for their viability as part of a last-ditch effort. Thus, there is a systematic bias in the sample toward successful prosecution experts, and appeals which are not necessarily based on a rational evaluation of the scientific merit of the testimony being challenged. These conditions make it unlikely that the results can be generalized meaningfully, or even taken at face value. Any effect on admissibility attributable to strength of the *Daubert* evaluation factors could easily have been obscured by excessive noise and restriction-of-range effects.

Beyond the problem created by the limits of the cases examined, there is a significant question as to whether the number of words devoted to discussion of a particular factor in a published decision represents a valid measure of the importance of that factor in the actual decision process. A wide variety of extraneous issues may affect word count. For example, a particularly complex issue related to general acceptance may be discussed at length, or extensive exposition on the court's understanding of falsifiability could be included in the decision, despite these factors

having little or nothing to do with the actual decision process. The relationship of word count and importance in the decision is ambiguous at best. Second, it cannot be assumed that the written text of a decision will accurately pinpoint the reason, or reasons, for the decision no matter how conscientiously judges attempt to explain their reasoning. The problem with decision-makers' potential lack of insight into their cue utilization and decision processes (described above) still remains.

It is of interest however, that of the expert testimony challenged in this sample, 40% of the experts were from the fields of social work, psychiatry, psychology, or other "social-behavioral scientists." This suggests, despite the shortcomings of the analysis, that social science testimony is subject to *Daubert* challenges with considerable frequency, and perhaps disproportionally so compared to the ratio of social scientists to experts in other fields.

Dixon and Gill (2001, 2002) analyzed trends in a random sample of 399 Federal District Court opinions from January 1980 through June 1999 that addressed challenges to expert testimony in civil cases. Judges also appeared to scrutinize the trustworthiness of proffered expert testimony with greater care following *Daubert*, based on the proportion of rulings that specifically addressed issues of scientific merit. At the same time, judges appeared to pay greater attention to relevance and expert credentials, which the authors attribute to increased focus on the gatekeeping role. Following *Daubert*, the proportion of evidence found to be untrustworthy, or insufficiently trustworthy, tended to increase over time, and the frequency of summary judgments[8] increased from 21% in the years preceding *Daubert* to 48% between July

[8] A summary judgment is a decision rendered by the court that disposes of a case in advance of trial, or before the conclusion of arguments and testimony, because no material issue of fact exists and one party

of 1995 and June of 1997. Although this change in the frequency of summary judgments is an interesting finding, there is little basis for assuming that it is due to the effects of *Daubert*. Other factors not addressed in the study could easily have been operating instead of, or in combination with, *Daubert* during the period under investigation and could account in whole or in part for the change in the frequency of summary judgments.

Dixon and Gill identified a total of 601 evidentiary elements in their sample of cases, which they sorted into 17 variables potentially related to the assessment of scientific merit. These included the *Daubert* criteria, as well as other scientific considerations (e.g., reliance on verifiable evidence or data) and extra-scientific considerations (e.g., expert's reputation; whether the research was done for the purposes of litigation). Of note, the only variables that were addressed in a greater proportion of post-*Daubert* cases with consistency were "general acceptance," the clarity and coherence of the expert's presentation, and reliance on verifiable evidence or data. Testimony from the social and behavioral sciences increased sharply in the proportion of evidentiary elements found to be inadmissible post-*Daubert*.

In summary, there has been little research on the effect of *Daubert* on admissibility decisions, the research that has been completed has methodological limitations that restrict interpretive value, and there has been no experimental research examining judges' actual decision-making process. However, it appears that social scientists may be a relatively likely target of challenge, and that they are excluded with

---

or the other is entitled to a judgment as a matter of law. For example, a party's argument over cause or liability may rest solely upon the testimony of an expert. If that expert were excluded, then that party would be unable to present evidence supporting an essential element of their claim, and therefore could not make their *prima facie* case. In this circumstance, the judge might dismiss the case.

some degree of frequency when challenges do occur. Thus, the subject is ripe for research but, so far, hardly investigated scientifically.

Actuarial versus Clinical Judgment

A discussion of how judges, attorneys, and testifying experts make appraisals of scientific merit requires a more general discussion of research in decision-making and clinical judgment. In particular, there is a 60-year tradition of research in clinical[9] versus actuarial judgment (Grove & Meehl, 1996; Meehl, 1954; Sarbin, 1943), and the findings remain among the most robust and consistent in the behavioral sciences (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Meehl, 1986). Actuarial judgment methods are explicitly prespecified, automated decision procedures based on empirically established relationships between variables. Examples of actuarial methods could include the determination of insurance rates from a table incorporating demographic and health status variables, or the selection of candidates for admission to college based on a regression formula including such information as high school GPA, class rank, and SAT scores. However, these decisions would only be actuarial if the prespecified procedure based on empirically established relations was followed and resulting decision outcome was, in fact, adhered to. That is, if an individual, based on his or her subjective judgment, made adjustments to the insurance cost after consulting the actuarial table as a guideline, or modified the rank-order of prospective candidates based on information not contained in the formula, then the decision

[9] For the purposes of this paper, 'clinical' refers broadly to all subjective judgment methods irrespective of the type of decision being made or the expertise or discipline of the decision-maker.

method is clinical rather than actuarial in nature, even though the individual utilizes actuarial results as data.[10]

Across over 200 direct comparisons, with equivalent data (or an informational advantage in favor of clinical judgment), actuarial methods nearly always equaled or exceeded the accuracy of subjective "in the head" methods of data combination (Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000). This finding holds true across a wide variety of classification and predictive tasks in the behavioral sciences and related fields (e.g., medicine, personnel evaluation). An actuarial decision procedure may incorporate both quantitative and qualitative data (including subjectively derived data such as clinical impressions), and such tools may be applied to virtually any decision process of interest (Dawes, Faust & Meehl, 1989).

Meehl (1992) has provided a detailed taxonomy of sources of error in clinical judgment, both motivated and non-motivated, that contribute to the relative superiority of actuarial methods. Among these are information overload, undue influence given to particularly salient cases, and more motivated biases based on theoretical or ideological influences. Actuarial formulae are perfectly reliable (in the sense that the same input will always produce the same output), make optimal or proper use of input data when designed correctly, and are not vulnerable to various errors related to human biases or the limited information processing capacity of the human mind.

Modeling Clinical Judgment

Although actuarial decision procedures are usually directed toward outcomes, one method of developing actuarial decision rules is the creation of statistical models

[10] Depending on the decision methodology, this could be described in some, but not all, cases as a "clinical-actuarial" approach (Dawes, Faust, & Meehl, 1989).

of human decisions[11]. For example, there may be cases in which a particular decision maker is recognized as accurate, or very accurate, when her decisions are compared to some objective standard. Assuming that a decision maker shows consistency in her judgments, and that relevant variables can be identified and controlled, one can empirically determine the statistical relationship between standing on background variables and her decision outcomes. Ample research shows that regression equations can be developed that will often predict an individual's judgments with modest to high levels of accuracy (Camerer, 1981; Dawes, 1979, 1986; Dawes & Corrigan, 1974; Goldberg, 1965, 1968, 1970, 1976; Hoffman, 1960; Wainer, 1976). Often, these equations need to make minimal use of possible interactions or configural relations among cues, even with judgment tasks that are claimed to be highly configural by decision-makers.

Furthermore, it is not sufficient to simply ask individuals how they go about making decisions, or how they use and combine information, because individuals' subjective estimates of the weights or influence of variables may be highly inaccurate, even if their judgments are, themselves, quite good. The way that individuals weigh and combine information must be examined statistically. For example, just as a regression equation can be used to predict an individual's decisions, it can also provide objective data on how those decisions are made. Research indicates that individuals' insight into the factors that influence their judgments and the weight attached to these factors is often quite poor: there may be minimal correspondence between decisive factors in decisions (as determined via statistical analysis) and the

[11] Models of human decision processes are sometimes referred to as "paramorphic" models in the judgment literature.

factors that individuals think were decisive in their decisions (Aspel, Willis, & Faust, 1998; Fisch, Hammond, Joyce, & O'Reilly, 1981; Gauron & Dickenson, 1966; Kirwan, Chaput de Saintonge, Joyce, & Currey; 1983; Nisbett & Wilson, 1977). This is not to say that individuals inevitably lack insight into the information they utilize in making decisions, but that formal study is necessary to determine the relationship between subjective impressions and measurable influences. That is, technology exists to study and compare subjective and objective cue utilization. These findings on decision-maker's insight align with research that demonstrates the limits of human information processing and the influence of various biases on human cognition (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Meehl, 1954).

To the extent that there is a systematic relationship between the information or cues available to a decision maker and the decisions made, then those cues may be used as predictor variables. The resultant formula will take into account redundancy between predictors, discount those variables that are non-contributory, and provide optimal weights for those variables with unique utility (although equal unit weighting, or in some cases even assigning weights randomly, may produce very similar outcomes and be comparable to the predictive decisions of human judges; Dawes, 1979).

In cases where outcomes are known, we may compare the accuracy of the statistical model of the expert to the expert on which that model is developed. To the extent that decision making procedures have validity, increasing the reliability or consistency of these procedures is likely to augment accuracy. Due likely in large part to the perfect consistency of the statistical model, the model's accuracy will often

exceed the accuracy of the individual upon which it is based. This phenomenon is referred to as the "Goldberg Paradox," and when it occurs, it can be said that we have "bootstrapped" our way to improved predictive power (Goldberg, 1970). The actuarial method for predicting decisions may then be used in place of the expert upon which it has been based, for example, in circumstances in which that individual is not available. As noted, reproducing the decision-making of individuals who are objectively accurate, or very accurate, makes their acumen broadly available, e.g., if Smith is an excellent diagnostician, one does not necessarily need to have Smith in one's clinic.

It is also sometimes of interest to predict decisions rather than outcomes. Parallel research suggests that actuarial methods also provide the best way to predict the decisions of individuals (or groups) when the decision in and of itself (independently of external evaluation of the decision's accuracy) is of interest (Dawes, 1986). For example, a major league baseball manager may wish to predict whether his opposite number will call a sacrifice bunt in a given situation, or an investor may wish to predict if the Chair of the Federal Reserve will adjust interest rates. Although one could certainly evaluate whether bunting or adjusting interest rates is a good move according to some external or objective standard, there are separate, and potentially significant, advantages to forecasting the decision itself with accuracy. For example, bunting may be a bad decision based on a statistical analysis of run production, but knowing that that one's opposition is going to do so allows one to set up a more effective defense, thereby providing a tactical advantage within the game. There may also be cases in which we lack access to outcomes but there is a reasonable basis to

conclude that decision makers (or at least some decision makers) may have "validity." The above-described research on the comparative accuracy of actuarial and clinical judgment applies in such circumstances, as we may forecast decisions via clinical judgment or through the use of actuarial means.

There are compelling reasons to believe that actuarial prediction should do as well as, or better than, attorneys utilizing their subjective judgment when predicting decisions within legal proceedings. When we wish to anticipate future decisions (e.g., attempting to forecast a judge's decision to admit or exclude expert witness testimony), an actuarial model of the decision-maker should be as accurate, or more accurate, than attempts to predict the decision through subjective ("in the head") means (Dawes, 1986). Furthermore, such a model would provide objective measures of the weight attached to different variables and a means to compare those weights to subjective impressions of the importance of those variables. The identification of discrepancies between subjective and statistically derived weights is of great potential relevance to legal scholars because if self-descriptions of the reasons for rulings do not comport closely to objective measures of influential factors, considerable legal history and writing will need to be examined and the role of "precedent" re-evaluated. Although the use of bootstrapped models of judges' decisions promises to provide insight into, and improved prediction of, admissibility rulings, and although it has been possible to develop such models in other domains, feasibility in the legal context has not been studied.

Modeling the Clinical Evaluation of Scientific Merit

When evaluating the quality of science for use in legal settings, the psychologist serving as an expert witness or consultant is likely to have an advantage over the average attorney or judge based on differences in training and experience. Training in research methods and statistics is a standard component of virtually all graduate programs in psychology, and may be the main focus within some specialties. Standard introductory textbooks from any core curriculum area in the field generally have at least one section or chapter to devoted research methods and the practice of accumulating knowledge through science. Even psychologists who espouse epistemologies that do not allow for knowledge (or even the existence) of an objective reality independent of our perceptions are conversant with the language and methods of scientific inquiry.

Despite this potential advantage in background knowledge, the psychologist is faced with a judgmental task that ultimately parallels that of the judge or attorney seeking to appraise the scientific foundations of expert opinion. *Daubert* has provided legal scholars and professionals with a list of factors to consider in evaluating the trustworthiness of scientific evidence, but, as noted above, the list is explicitly non-exhaustive, non-ordinal, and non-hierarchical, and no guidance was provided as to the relative weight to be placed on individual factors or how to resolve inconsistencies or contradictions should they arise. Inconsistencies or contradictions do arise in most non-trivial cases, or cases that pose any difficulty. That is, if all indicators point unanimously toward a single conclusion, then the evaluation of scientific merit is a

relatively simple task. This is not typically the case in the social sciences (Meehl, 1990).

Psychologists, due to their training and experience, might well identify a different list of factors that they would consider important when evaluating the trustworthiness of the same evidence including, in addition to, or in place of, those that *Daubert* specified.[12] However, once the psychologist has identified this list of factors, there is no standard, whether logically derived or empirically validated, to assist her in determining the optimal way to utilize this list (for example, as would be available if the objective relationship between scientific factors used in theory appraisal and the scientific status of theories were, in fact, know). As is the case with the list of factors that *Daubert* provided to the legal community, the psychologist's list is non-exhaustive, non-ordinal, and non-hierarchical, and there is no ultimate authority that has established the relative weight to be placed on individual factors or how to resolve inconsistencies or contradictions should they arise (c.f., Faust, 1997; Faust & Meehl, 1992, 2002; Meehl, 2002).

For example, a particular study might have the advantages of a sufficient sample size, representative sampling methods, a plausible theoretical model, parsimony of explanation, and provide rigorous statistical tests of the hypotheses under investigation. Although each of these factors can be argued to be positive, none will be perfectly predictive of the ultimate trustworthiness of the study. Further, although it may be the case that some flaws could be seen as conclusive in dismissing

[12] For example, Meehl (2002) identifies eighteen factors that at least some scientists take into account at least some of the time when appraising scientific theories. Although he was not writing about theory appraisal in the legal context, the general point that scientists may vary in their use of a large number of factors remains relevant to the problem under consideration in this dissertation.

a study (for example, an extreme and obvious sampling bias), the standing of studies on most factors will likely be more ambiguous. How, then, is the psychologist to evaluate the merit of a study in which the sample size and sampling methods are excellent, but the measures used are below average? How is the psychologist to determine the relative merit of two studies that have generated inconsistent results and present different combinations of strengths and weaknesses from an evaluative standpoint?

Although it is certainly possible that the factors scientists identify for evaluating science have stronger associations with objective scientific standing than factors the court identifies, that is an empirical question, and a minor one in an important regard: In the absence of objective knowledge of the relationship between identifiable characteristics of scientific studies, theories, or programs of inquiry (evaluative factors or cues) and the true scientific status of those studies, theories, or programs of inquiry, the psychologist, attorney, and judge are all in the same position. Each must make a subjective judgment under conditions of uncertainty utilizing probabilistic indicators, and each decision is therefore subject to all of the same biases and limitations affecting clinical judgment. This fundamental problem of the evaluation of science in the face of our human limitations is indeed the rationale for the field of cliometric metatheory or meta-science, which seeks to examine problems in the history and philosophy of science through the application of scientific methods (Faust, 1984, 1997; Faust & Meehl, 1992, 2002; Meehl 1992, 2002, 2004).

The Current Study

The preceding review integrates the literature from a number of areas spanning, law, the methods and philosophy of science, and research on judgment and decision-making. The goal of this dissertation was to lay the groundwork for a program of research on the decision-making of judges, attorneys, and expert witnesses on tasks related to the appraisal of the scientific merit of expert witness testimony. The scope and complexity of the problem suggests many avenues of research, and a myriad of potential studies. The present study is a small initial step, designed as a point of entry into this rich territory.

The author developed a survey based on review of the literature on *Daubert* and the evaluation of scientific evidence, a review of *Daubert* cases involving social science testimony, and interviews with members of the Rhode Island judiciary. The survey was administered to forensic psychologists, who were asked to provide information about how much weight they place on a series of factors potentially related to the evaluation of scientific merit. They were also asked to estimate how much weight they believed judges would place on a subset of these factors.

Forensic psychologists were selected as the population for study for a number of theoretical and pragmatic reasons. First and foremost, psychologists generally have more training in science and the scientific method than do judges and attorneys. It is arguably sensible to begin the study of the scientific evaluation of courtroom testimony with individuals who have scientific training in an effort to identify variables that may have a stronger relationship to true scientific standing. The selection of this population also allows for the methodology of the study to be tested

and refined prior to moving forward with surveys of judges and attorneys. These populations are notoriously difficult to gain access to for research purposes, and it would be prudent to ensure that the methods used in future studies have the best chance to produce meaningful results.

Based on this survey, the primary aims of the study were to:

1) Determine the self-reported weights assigned by psychologists to factors of potential relevance to evaluating the scientific trustworthiness of expert witness testimony in the legal setting, i.e., to determine psychologists' subjective estimates of cue utilization when appraising scientific merit for courtroom use.
2) Evaluate the degree to which psychologists agree (or do not) as a group on their self-reported weights assigned to these factors of interest.
3) Determine the impressions psychologists have of the weights judges would assign to a subset of these to factors, i.e., to determine psychologists' estimates of judges' cue utilization when appraising scientific merit for courtroom use.
4) Evaluate the degree to which psychologists agree (or do not) as a group on their estimates of the weights judges would assign to these factors.

Depending on the outcome of the above primary aims, the secondary and exploratory aims of the study were to:

5) Determine what differences, if any, can be identified between psychologists with different training and different experience with the legal system in the self-reported weights assigned to these factors.

6) Determine what differences, if any, can be identified between psychologists with different training and different experience with the legal system in their estimates of the weights judges assign to these factors.

7) Determine what differences, if any, can be identified between psychologists' self-reported weights and their estimates of weights assigned by judges to these factors.

Depending on the outcome of the above primary aims, and presuming a sample size in excess of N = 200, an additional exploratory aim of the study was to:

8) Examine the intercorrelation of participants' self-reported subjective weights.

In overview, and in lines with the above-stated aims, examination of the extant literature, a review of *Daubert* decisions on social science testimony, and semi-structured interviews with six members of the Rhode Island judiciary were used to develop a list of factors that might be considered when evaluating the trustworthiness or merit of science for legal purposes. Participants' subjective estimates of the importance they would place on a set of these factors were measured by having the participants rate the factors on a five-point Likert scale. Participants' estimates of the importance they believe a judge would place on a subset of these factors was measured by having the participants rate the weight they believe judges would place on these factors using the same five-point Likert scale.

It should be noted that this study was designed to examine consistency in estimated weights, and possible differences in estimated weights between groups. The "accuracy" of participants' appraisals as defined by an external standard was not the focus of this research, but rather the decision process itself.

Consistency in the rulings of judges and their appraisals of scientific trustworthiness[13] is of great practical importance, as the outcome of legal proceedings should arguably be determined by the facts at issue and the relevant law. Although judges do have latitude in the interpretation and application of various elements of the law, the fate of the parties to a case should not be principally, or largely, determined by which judge they happen to draw, or by luck. It is hard to conceive of chance as enhancing a rational or fair system of justice or ethics. As judges and attorneys may be heavily reliant upon the appraisals of expert witnesses and consultants on matters of scientific merit, investigation into the bases of these appraisals is an important first step in disentangling the various elements that result in the ultimate admission or exclusion of expert witness testimony at trial.

Further, it is to the benefit of all parties involved in litigation to gain insight into their appraisals of scientific merit, and to identify discrepancies in the cues considered important by individuals with different roles in legal proceedings. The examination of the decision process itself is an important step in response to changes in the legal code and the role of the judges as gatekeepers, and it is hoped that legal and scholarly attention to how such decisions are made, and how they should be made, will promote just legal outcomes.

Finally, although this study can only serve as a starting point for the following aim, developing actuarial methods for evaluating the likelihood that a *Daubert* challenge will succeed could have salutary effects for the Court, practicing attorneys,

[13] As noted previously, the terms "validity" and "reliability" have different meanings in psychological methodology and in the legal system. As the testifying psychologist enters the legal context as an invited guest, it is appropriate to use the language common to legal institutions, scholarship, and actors. Therefore, "trustworthiness" or "merit" will be used when discussing the validity of methods.

their clients, and the expert witnesses and consultants practicing in the *Daubert* era. A more accurate evaluation of the scientific merit of potential testimony and its likelihood of admissibility should improve legal proceedings by reducing spurious *Daubert* challenges, increasing the probability that junk science can be identified and excluded, and bolstering the quality of experts called to testify on matters of science[14].

METHOD

Participants

Names and mailing addresses for a random sample of N = 600 psychologists were obtained through the American Psychological Association's Center for Psychology Workforce Analysis and Research. Only full members with a Ph.D. or its equivalent, and who are residents of the United States, were included in this sample. Membership in APA Division 41 (The American Psychology-Law Society) and/or listing of forensic psychology as a primary area of interest in the APA membership directory was also an inclusionary criteria. These criteria were selected in consultation with the Center for Workforce Analysis and Research in an effort to target individuals with relevant professional training, education, and experience in forensic work, and who would be likely to have experience with courtroom testimony as an expert witness.

[14] For example, such knowledge could assist attorneys in screening potential experts in the early stages of trial preparation, increasing the likelihood that they can identify experts who employ scientifically trustworthy methods. Such experts may be better able to assist attorneys with sound consultation at each subsequent stage of the litigation process and should be better able to withstand admissibility challenges.

Procedure

Preliminary interviews. Based on contact with the Rhode Island Superior Court, judges with an interest in admissibility issues were interviewed on a volunteer basis following a semi-structured format. Six judges were included in the preliminary interview process, for 1 to 2 hours each. As this is a novel area of research, the interview was modified iteratively based on the feedback received. Additional consultation has been provided by a Federal Magistrate with expertise in Daubert and the admissibility of expert witness testimony. The goal of the interview process was not data collection in itself, but to gain insight into the view of *Daubert* from the bench, how admissibility issues are decided, and to solicit suggestions for areas of inquiry, potential variables, and the format and presentation of surveys and future hypothetical case scenarios.

It should be noted that the *Daubert* criteria are somewhat vague or open to interpretation, and perhaps intentionally so in recognition of the importance of judicial discretion in applying the law to the facts of a given case at issue. Beyond the issue of the weight judges may place on these factors, judges may disagree as to the minimum threshold that each factor must meet in order for testimony to be admitted. For example, as noted above, the *Daubert* decision references the "known" error rate of methods or procedures. There is debate in legal scholarship as to whether this should be taken as meaning that the judge should establish a maximum error rate he or she would consider acceptable (a standard which could vary by scientific field, whether the case at hand is a criminal or civil matter, or various facts specific to the case) in order for the testimony to be admitted, or whether this simply means that a method

with an error rate of any magnitude should be admitted provided that the error rate is known. It has even been argued that a method with an error rate known to be 100% should be admitted by the judge, with the magnitude of that error rate to be properly addressed during cross-examination and going to the weight that the trier of fact ascribes to the testimony. At issue is not the ultimate reliability of the testimony, as no reasonable person would argue that a method that is always wrong could meet any meaningful standard of trustworthiness; it is the limits of judicial discretion and the interpretation of the Supreme Court's intent that is the subject of debate. This is likely to be a purely academic point. Were an attorney to successfully gain the admission of an expert witness whose testimony was based on methods wrong all, or nearly all, of the time, it would almost certainly be a Pyrrhic victory.

Beyond these pragmatic concerns, there is a larger issue that lies in the different language often used by judges and scientists when discussing the same phenomena. In the design of this study generally, and in the writing of specific questionnaire items in particular, the author was been in the difficult position of trying to understand the view from the bench, which varies from judge to judge and even case to case, and then translating the language used to express that view into terms more readily understood by psychologists who vary greatly in their legal training and experience. Although no wording of items will capture the true meaning of these often complex and ambiguously defined concepts exactly, the input from these esteemed members of the bench was invaluable in attempting to represent a consensus view that can be studied with some degree of scientific validity.[15]

[15] In this case, there was no unanimous agreement on any issue, and so "consensus" refers only to a significant degree of shared understanding on specific issues among some individuals. While there was

Survey Design. In the absence of an existing survey suited to the goals of this study, the following method was employed for the design of an appropriate instrument. Variables that may influence the outcome of a *Daubert* challenge were identified by a relatively exhaustive process: a review of the literature, a review of *Daubert* cases involving social science testimony, and interviews with judges as described above. In addition to these sources specific to *Daubert* and admissibility issues, literature was also reviewed that provided information on heuristics for the evaluation of scientific evidence (e.g., Bradford-Hill, 1966; Meehl, 2002; Meltzoff, 1998; National Academy of Sciences, 2000; Park, 2003; Ruscio, 2006; Sagan, 1996) and the philosophy of science (e.g., Buchdahl, 1970; Cohen & Nagel, 1934; Faust, 1984, 1997; Faust & Meehl, 1992, 2002; Feyerabend, 1993; Hempel, 1966; Kuhn, 1968; Margenau, 1950; Meehl, 1992, 2002, 2004; Newton-Smith, 1981; Popper, 1959, 1963; Rescher, 1990; Salmon, 1998; Schaffner, 1970; Shapere, 1977; Thagard, 1978, 1992).

This review identified factors specified in *Daubert*; factors specified in subsequent court decisions that followed, referenced, or relied upon *Daubert* or that appeared in legal scholarship analyzing *Daubert*; various characteristics of scientific studies or programs of research; and factors not related to scientific trustworthiness *per se*, but still considered relevant to the admissibility of expert witness testimony. During this process, the author identified sufficient evidence to hypothesize that judges place significant weight on their evaluation of the witness's credibility, which may be related to the strength of the underlying testimony but may also include factors

---

certainly individual variation, the emergence of common themes and ideas was interesting, informative, and of practical value in the design of the study.

such as the expert's personality, credentials, presentation, or the judge's impression that the expert is honest or unbiased[16].

This raises an interesting issue with regard to how the appraisal of scientific trustworthiness actually relates to the admissibility decisions of judges: it is possible that judges may invoke, or make reference to, *Daubert* and scientific merit when explaining their exclusion of an expert in circumstances when their gut impression is that the expert is not trustworthy for other reasons. That is, *Daubert* may provide a basis for justifying the exclusion of an expert who fails the credibility test alone, and therefore serves as a means to an end.

Following the investigations described above, variables bearing directly on scientific reliability (including, but limited to, those specified in *Daubert*), and extra-scientific variables, such as expert credentials and credibility, were identified for possible inclusion in the survey. Working from this general list of variables, a subset was selected for inclusion in the survey. Some selectivity in determining which variables to include was necessary and in some sense desirable, but it was not clear, absent previous research, which variables should be studied or were most important. Consequently, the choice of certain variables over others was somewhat arbitrary. Potential variables were evaluated based on the following considerations:

1. Amenability to objective, or relatively non-ambiguous, written description, e.g., it is much easier to describe the presence or absence of a certain

[16] For example, in an early discussion of the factors related to the admission of expert witnesses, one individual generated a list of factors that he believed, based on his considerable experience, had a significant influence on judges' admissibility decisions. Of the twenty-plus factors he initially identified, not one related to actual scientific status; issues related to expert credibility were predominant. It was clear from this discussion that he did not believe that such decisions were arbitrary, or that scientific merit was irrelevant to the decision-making of judges; however it is also clear that credibility is perceived by judges and others to be an important consideration.

credential, such as a Ph.D., as opposed to a more nebulous quality, such as the degree to which a theory is amenable to risky tests of falsification.

2. Frequency of mention or particular emphasis by judges during interviews.
3. Frequency of mention in literature, as regular reference to specific factors (e.g., the error rate of techniques) in analysis or in published court decisions may reflect the subjective impression that this factor is given significant weight. It also seems sensible to focus on variables courts seem to deem as most, or among the most, important.
4. Qualitative diversity, as there is widespread differences of opinion expressed in discussion and in the literature about what is (most) relevant, and the extent to which various proposed variables do, or should, influence admissibility decisions. For example, in an early discussion of the project with a prominent attorney that covered variables he thought might be most influential, almost every one he listed was a non-scientific consideration. Therefore, issues such as the perceived credibility of expert witnesses may have considerable weight, and would be feasible to include as variables.
5. Scholarly interest. For example, the inclusion of "general acceptance" is clearly warranted as should it be determined that the *Frye* standard is a main factor in determining admissibility then the practical relevance of *Daubert* becomes questionable.

Data collection. Participants were mailed a packet containing an introductory letter, informed consent document, survey, postage-paid return envelope for the survey, and postage-paid postcard to be used to request a copy of the results once the study was

completed. A follow-up reminder card was mailed two weeks after the initial packet. These items are presented in Appendices A through E.

The survey asked participants to provide basic information related to their training and education, information related to their experience with courtroom testimony and consultation, and their opinions on some issues related to admissibility.

Participants were asked to indicate the relative weight they believe they would place on a list of general factors of potential relevance to evaluating scientific merit using a five-point Likert scale ranging from "No Weight" to "Great Weight." Examples of these factors include the error rate of the method and whether or not the method is generally accepted within the relevant scientific community.

Participants were next asked to indicate the relative weight they estimate a judge would place on the same set of general factors using the same five-point Likert scale ranging from "No Weight" to "Great Weight."

Subsequently, participants were asked to indicate the relative weight they believe they would place on a list of more specific factors for evaluating scientific merit using the same five-point Likert scale ranging from "No Weight" to "Great Weight." Examples of these specific factors include "The method is accepted by most, but far from all, members in its field," and "The method has an error rate of 20%."

Finally, participants who indicated that they had provided expert witness services as part of legal proceedings were asked to answer two additional questions about their personal experience with challenges to their admissibility.

After completing the survey, participants had the opportunity to provide an address to receive a summary of the results once the study was completed. They were reminded that their participation in the study is anonymous and confidential, and that their individual responses could not be linked to their identity should they choose to request a summary of the results.

Data Analysis.

This section will begin with a review of the characteristics of the obtained sample, followed by the presentation of participants' responses on several questions related to minimum standards for error rates and general acceptance, and then a summary of results related to the specific research goals of the dissertation.

139 participants returned surveys; of these, 13 returned only partially completed surveys and were removed from analysis, leaving 126 surveys for analysis. The return rate of 21% is relatively low, and it is not possible, given the lack of demographic data collected, to analyze factors that could differentiate responders from non-responders. The representativeness, or lack thereof, of the sample is therefore uncertain, a point to be taken up in more detail later.

The obtained participant pool was relatively homogeneous across various dimensions. The modal participant held a Ph.D. (79%), did not graduate from a program specializing in psychology and law (91%), and had completed 0 – 1 full semester courses on the evaluation of scientific evidence for the courtroom (73%). Participants reported a wider range in the number of brief professional or continuing education workshops or seminars on evaluating the trustworthiness or merit of scientific evidence in the legal setting completed. As can be seen in Table 1, although

Table 1

Brief Professional or Continuing Education Workshops or Seminars on Evaluating the Trustworthiness or Merit of Scientific Evidence in the Legal Setting Completed

Number of Courses	Frequency	Percentage	Cumulative Percentage[17]
0	10	8%	8%
1 – 10	62	50%	58%
11 – 20	15	12%	70%
21 – 30	11	8%	78%
>30	26	21%	99%

[17] Percentages may not add to 100% due to rounding. This applies to all tables.

ten participants reported no continuing professional education on the evaluation of scientific evidence in legal settings, fully half indicated some such training, and the remainder reported that they had completed eleven or more courses or seminars of this type. As Table 2 shows, roughly half of participants reported that they spent 50% or more of their professional time engaged in forensic work directly related to courtroom activities including, for example, preparation, assessments, report-writing, consultation and testimony.

The majority of participants reported that they had been proffered as an expert witness in one or more criminal and/or civil cases within the past 10 years (see Table 3). Although *Daubert v Merrill Dow* was decided in 1993, the effects of the ruling on how psychologists evaluate evidence for the courtroom would not have been instantaneous. Collecting data on this time-frame was intended to allow for some lag in the diffusion of knowledge and adaptation to the new rules. As shown in Table 4, the majority of participants indicated that they are at least somewhat familiar with the guidelines for evaluating the trustworthiness of proposed scientific evidence under *Daubert.*

These results indicate that the majority of participants have some, or even considerable, experience with courtroom-related professional activities, and that the majority have been proffered as expert witnesses. In addition, about 80% of the respondents report at least moderate familiarity with the standards for the admissibility of expert witness testimony under *Daubert v. Merrill Dow*. The obtained sample is therefore consistent with the target population for the study in that participants are generally qualified to answer questions about the merits of scientific testimony in the

Table 2

Percentage of Participants' Professional Time Spent on Forensic Work Directly Related to Courtroom Activities.

Time Spent	Frequency	Percentage	Cumulative Percentage
≤10	26	22%	22%
20	15	12%	34%
30	12	10%	44%
40	7	6%	50%
50	12	10%	60%
60	8	6%	66%
70	10	8%	74%
80	5	4%	78%
≥90	31	24%	100%

Table 3

Frequency with which Participants Were Offered as Expert Witnesses within the Past Ten Years.

# Cases	Criminal Court (Percentage)		Civil Court (Percentage)	
0	18	(14%)	28	(22%)
1 – 5	24	(19%)	19	(15%)
6 – 10	12	(10%)	9	(7%)
10 – 15	7	(6%)	8	(6%)
>15	64	(51%)	61	(49%)

Table 4

Participants' Familiarity with Daubert v. Merrell Dow.

Familiarity	Frequency	Percentage	Cumulative Percentage
Not Familiar	3	2%	2%
Somewhat Familiar	21	17%	19%
Moderately Familiar	55	44%	63%
Very Familiar	46	37%	100%

legal context, and have sufficient experience in the legal domain to consider how they and judges may evaluate scientific evidence. As noted above, the relatively low return rate raises the issue of some unknown, but systematic, biasing factor that distinguishes responders from non-responders. There is no way to directly evaluate this possibility, calling for caution in attempts to generalize the results to a broader population. However, given the paucity of research in this area, the characteristics of the sample that was obtained, and the exploratory nature of this dissertation, these limitations would not seem to obviate the potential value of the work.

In addition to these background variables, participants were asked several questions related to two of the four main evaluation criteria specified in Daubert: error rate and general acceptance. As discussed above, although *Daubert* specified the "known or potential error rate" of a scientific technique as one of the criteria for evaluating the trustworthiness of expert testimony, it provided no specific guidance on the subject of what an acceptable rate of error might be. Participants were asked to specify the maximum individual error rate (expressed as a percentage of errors) for any one test or method that they consider acceptable for use in at trial in both civil and criminal venues. They were also asked to specify the maximum cumulative error rate for a group of tests or methods taken as a whole that they consider acceptable for use in at trial in both civil and criminal venues.

These results, which appear in Table 5, indicate that most participants endorse an error rate of no more than 20% for any single method to be introduced as part of expert witness testimony at either a civil or criminal trial. There is a slight difference between the minimum acceptable error rate for civil trial as compared to criminal trial,

Table 5

Participants' Maximum Acceptable Error Rate for Methods to be Used at Trial

	Civil Trial		Criminal Trial	
Errors	Individual Rate	Cumulative Rate	Individual Rate	Cumulative Rate
≤ 10%	48 (44%)	43 (40%)	63 (56%)	56 (50%)
20%	30 (28%)	38 (34%)	26 (23%)	33 (30%)
30%	21 (19%)	16 (14%)	13 (12%)	11 (10%)
40%	6 (6%)	10 (9%)	7 (6%)	7 (6%)
50%	1 (1%)	0 (0%)	0 (0%)	1 (1%)
60%	1 (1%)	1 (1%)	1 (.1%)	1 (1%)
70%	1 (1%)	0 (0%)	0 (0%)	0 (0%)
80%	1 (1%)	1 (2%)	2 (2%)	2 (2%)
≥ 90%	0 (0%)	0 (0%)	0 (0%)	0 (0%)

with 72% of the participants endorsing the 20% threshold in a civil context and 79% endorsing the 20% threshold in a criminal context. Participants endorsed a cumulative error rate, when considering a body of procedures in combination, that was roughly equivalent to their minimum standards for methods considered individually: 74% of the participants endorsed a cumulative error rate of 20% or less for admissibility in a civil trial, and 80% endorsed that threshold for admissibility in a criminal trial. These differences were not statistically significant.

In addition to the questions on error rate, participants were asked two questions relevant to the Frye standard of general acceptance:

> What percentage of agreement within a professional field surpasses the threshold for a method to be considered generally accepted in that field?
>
> What percentage of agreement within a professional field surpasses the threshold for a method to be considered accepted by a sizeable minority in that field?

As seen in Table 6, participants' responses spanned the full range of response options. Participants offered a minimum threshold for a method to be "generally accepted" as low as 10% or less, and roughly 20% of the participants set the threshold at 40% or less. Likewise, participants offered a minimum threshold for a method to be accepted by a "sizeable minority" as high as 90% or more, and roughly 34% set the minimum threshold at greater than 50%. Conceptually, it is difficult to understand how a method can be considered "generally accepted" when it is endorsed by a minority within the field, or how a "sizeable minority" can be defined as exceeding 50% of the field.

Table 6

Participants' Thresholds for "General Acceptance" and a "Sizeable Minority."

	General Acceptance		Sizeable Minority	
	Frequency	Cumulative %	Frequency	Cumulative %
≤10 %	4	3.4 %	4	3.6 %
20 %	9	11.2 %	19	20.9 %
25 %	1	12.1 %	0	20.9 %
30 %	5	16.4 %	29	47.3 %
40 %	5	20.7 %	14	60.0 %
50 %	8	27.6 %	7	66.4 %
60 %	12	37.9 %	8	73.6 %
70 %	39	71.6 %	13	85.5 %
75 %	1	72.3 %	0	85.5 %
80 %	19	88.8 %	11	95.5 %
85 %	1	89.6 %	0	95.5 %
≥90 %	12	100 %	5	100 %
Median		70.0 %		40.0 %
Mode		70.0 %		30.0%

Also notable is the frequency of extreme values endorsed by some participants: 10% of participants set the threshold for general acceptance at 90% of the field, a standard unlikely to be met by most theories and methods within the field. In contrast, nearly 4% of participants endorsed 10% or less as the threshold for a sizeable minority, a standard that could be met by a wide range of theories and methods irrespective of their underlying scientific merit.

Research Goal 1

Determine the self-reported weights assigned by psychologists to factors of potential relevance to evaluating the scientific trustworthiness of expert witness testimony in the legal setting, i.e., to determine psychologists' subjective estimates of cue utilization when appraising scientific merit for courtroom use.

The survey asked participants to indicate how much weight they would place on 32 items on a five-point Likert scale: "1 = No Weight," "2 = Little Weight," "3 = Some Weight," "4 = Moderate Weight," and "5 = Great Weight." The items included general considerations, such as the error rate of the method, and more specific ones, such as an error rate of 20%, 50%, or 80%. Examination of Table 7 reveals that participants reported that they place moderate to great weight on all variables, with the exception of face validity (which was still accorded "some" weight).

Research Goal 2

Evaluate the degree to which psychologists agree (or do not) as a group on their self-reported weights assigned to these factors of interest.

The absolute agreement of weights assigned by participants to variables was assessed by calculating Cronbach's alpha coefficient. Participants were in striking

Table 7

Participants' Self-Reported Weights

Item[18]	Mean (SD)	Median	Mode
Whether or not the method is generally accepted within the relevant scientific community.	4.52 (.68)	5	5
The error rate of the method.	4.15 (.79)	4	4
Whether or not the method can be tested	4.37 (.79)	5	5
Whether or not the method has been tested	4.40 (.73)	5	5
Studies of the method have been published in peer-reviewed sources.	4.46 (.68)	5	5
The presence of, and conformity to, standards for the administration and interpretation of the method.	4.67 (.56)	5	5
Whether or not the expert has given due consideration to viable alternative explanations for the results obtained by the method.	4.51 (.66)	5	5
Whether or not the testimony appears to be based on sufficient facts or data.	4.86 (.34)	5	5

[18] Items are presented in the same order that they appeared in the survey.

Table 7, *cont.*

Participants' Self-Reported Weights

Item	Mean (SD)	Median	Mode
Whether or not the expert has followed standards of practice adopted in his or her field.	4.67 (.56)	5	5
The degree to which the method has face validity.	3.15 (1.03)	3	3
The degree to which the testimony employs parsimony of explanation.	3.46 (.90)	4	4
The degree to which the method has construct validity.	4.17 (.73)	4	4
The method is generally accepted in its field.	4.52 (.68)	5	5
The method is generally not accepted by most members in its field.	4.15 (.78)	4	4
The method is accepted by most, but far from all, members in its field.	4.37 (.79)	5	5
The method is rejected by most, but far from all, members in its field.	4.40 (.73)	5	5
The method is neither clearly accepted nor rejected in its field.	4.46 (.68)	5	5

Table 7, *cont.*

Participants' Self-Reported Weights

Item	Mean (SD)	Median	Mode
The method is generally accepted by practitioners (i.e., psychologists providing assessment and treatment) but research studies are generally negative.	4.67 (.56)	5	5
The method is generally not accepted by practitioners (i.e., psychologists providing assessment and treatment) but research studies are generally positive.	4.51 (.66)	5	5
There is no way for the method's error rate to be determined.	4.86 (3.4)	5	5
The error rate of the method could be determined but it has not been.	4.68 (.64)	5	5
The method has an error rate of ≤ 20.	3.15 (1.03)	3	3
The method has an error rate of 30%.	3.46 (.90)	4	4
The method has an error rate 50%.	4.17 (.73)	4	4
The method has an error rate 70%.	3.49 (1.74)	4.5	5
The method has an error rate of $\geq 80\%$.	3.49 (1.9)	5	5

Table 7, *cont.*

Participants' Self-Reported Weights

Item	Mean (SD)	Median	Mode
The method does not have standard administration procedures.	3.73 (1.4)	4	5
The expert has followed standard administration procedures for the method.	4.53 (.60)	5	5
The expert has made minor violations of standard administration procedures for the method	3.51 (.96)	4	4
The expert has made significant violations of standard administration procedures for the method.	4.20 (1.38)	5	5
The expert has made total and gross violations of standard administration procedures for the method.	4.35 (1.49)	5	5
The expert did not retain the raw data underlying his or her opinions.	4.29 (1.33)	5	5

agreement in their impression of the subjective weights that they place on the variables they were asked to consider ($\alpha = .84$), and reported that they give significant consideration to all of the variables identified. There is a consensus among participants that all variables should be accorded at least some, and perhaps significant weight when evaluating the scientific merit of testimony.

Research Goal 3

Determine the impressions psychologists have of the weights judges would assign to a subset of these to factors, i.e., to determine psychologists' estimates of the cue utilization of judges when appraising scientific merit for courtroom use.

The third goal of this study was to determine psychologists' estimates of the weights judges place on factors of potential relevance when evaluating the scientific trustworthiness of expert witness testimony in the legal setting, i.e., to determine psychologists' estimates of judges' cue utilization. The survey asked participants to rate 12 items on five-point Likert scale: "1 = No Weight," "2 = Little Weight," "3 = Some Weight," "4 = Moderate Weight," and "5 = Great Weight." These items were identical to 12 of the items that participants had given subjective weights for. Thus, participants provided both their subjective weights and their estimates weights for judges for these items. Overall, as seen in Table 8, participants estimated that judges place some weight to great weight on all variables. The comparison of participants' subjective weights of cue utilization and their estimated weights of the cue utilization of judges is discussed below under Research Goal 7.

Table 8

Participants' Estimated Weights for Judges

Item[19]	Mean (SD)	Median	Mode
Whether or not the method is generally accepted within the relevant scientific community.	4.29 (.72)	5	5
The error rate of the method.	3.07 (1.02)	3	3
Whether or not the method can be tested	3.37 (1.00)	3	3
Whether or not the method has been tested	3.63 (1.00)	4	4
Studies of the method have been published in peer-reviewed sources.	3.86 (.94)	4	4
The presence of, and conformity to, standards for the administration and interpretation of the method.	3.74 (1.03)	4	3
Whether or not the expert has given due consideration to viable alternative explanations for the results obtained by the method.	3.67 (.96)	4	4

[19] Items are presented in the same order they appeared on the survey.

Table 8, *cont.*

Participants' Estimated Weights for Judges

Item	Mean (SD)	Median	Mode
Whether or not the testimony appears to be based on sufficient facts or data.	4.36 (.82)	5	5
Whether or not the expert has followed standards of practice adopted in his or her field.	4.10 (1.01)	4	5
The degree to which the method has face validity.	3.71 (1.14)	4	4
The degree to which the testimony employs parsimony of explanation.	3.36 (1.14)	3	3
The degree to which the method has construct validity.	2.91 (1.03)	3	3

Research Goal 4

Evaluate the degree to which psychologists agree (or do not) as a group on their estimates of the weights judges would assign to these factors.

The absolute agreement of weights assigned by participants to variables was assessed by calculating Cronbach's alpha coefficient ($\alpha = .84$). Participants were in agreement in their estimates of the weights that judges place on these variables when considering the scientific merit of proposed testimony.

Research Goal 5

Determine what differences, if any, can be identified between psychologists with different training and different experience with the legal system in the self-reported weights assigned to these factors.

Although the obtained sample was homogeneous on various dimensions, several comparison groups could be distinguished based on post hoc analysis of the frequency distribution of variables. The following variables were selected as grouping variables for comparison: the number of full semester courses in psychology and law taken (0 compared to 1 or more); the number of continuing education seminars or workshops on the evaluation of scientific evidence for courtroom use (10 or fewer compared to 11 or more); the percentage of professional time spent on court-related forensic activities (50% or more compared to less than 50%); the number times the psychologist has testified in criminal cases (15 or more compared to less than 15); the number of times the psychologist has testified in a civil case (15 or more compared to less than 15); and participants' primary state of employment (states following the

Daubert standard, states following the Frye standard, and states following a mixed or hybrid standard).

There is a logical basis for one of these grouping variables that was anticipated in advance: participants' state of primary employment. State courts and legislatures have varied in their reaction to the *Daubert* decision. Some states have adopted standards for admissibility that are consistent with the standards established by the *Daubert* Trilogy (i.e., *Daubert v Merrell Dow, Kumho v Carmichael, and General Electric v Joiner).* Other states have rejected the Trilogy entirely and continue to follow the *Frye* standard (i.e., general acceptance). Still others have adopted hybrid standards (e.g., one standard for criminal cases and another for civil cases) or have adopted the standards of the original ruling in *Daubert v Merrill Dow*, but only one (or neither) of the modifications established by the other elements of the Daubert Trilogy (Bernstein & Jackson, 2004). As psychologists working primarily in states with different admissibility standards may reasonably be hypothesized to have had different experiences with the evaluation of the scientific merit of testimony during the legal process, the division of the sample based on the standard that prevails in their state of primary employment makes some sense.

The remaining five grouping variables were identified based on approximate median splits. The use of crude median splits based on post hoc analysis of the frequency distributions is not a strong method for hypothesis testing but, given the exploratory nature of this dissertation and the absence of prior research in this area, it seemed reasonable to proceed with these comparisons while acknowledging the limitations placed on the results obtained.

Independent t-tests (Bonferroni corrected) were calculated on psychologists subjective weights for 32 variables, comparing participants who have completed one or more full semester courses in psychology in law ($\underline{n} = 51$) and those who have not completed any full semester courses in psychology and law ($\underline{n} = 73$); comparing participants who have completed 10 or fewer continuing education seminars or workshops on the evaluation of scientific evidence for courtroom use ($\underline{n} = 52$) and those who have completed 11 or more ($\underline{n} = 72$); comparing participants who spend 50% or more of their professional time on court-related forensic activities ($\underline{n} = 55$) and those who spend less than 50% of their time on such activities ($\underline{n} = 69$); comparing participants who have testified in 15 or more criminal cases ($\underline{n} = 61$) and those who have testified in fewer than 15 ($\underline{n} = 63$); comparing participants who have testified in 15 or more civil cases ($\underline{n} = 61$) and those who have testified in fewer than 15 ($\underline{n} = 64$). Out of all of these comparisons, only significant three significant differences were obtained. As seen in Table 9, the number of continuing education seminars and workshops on the evaluation of scientific evidence participants have attended (≤ 10 versus $11 \geq$) was the grouping variable related to differences in the subjective weights participants assigned to three variables.

Participants with more continuing education on the evaluation of scientific evidence for courtroom use placed greater self-reported weight on each of these three variables. Even with Bonferroni correction for Type I error, the probability that one or more of these significant results is due to chance obviates meaningful interpretation of these differences. Retrospective analysis indicated statistical power of .79, roughly

Table 9

Significant Differences Between Participants' Self-Reported Weights for Variables used to Evaluate Scientific Merit based on Quantity of Continuing Education.

	≤ 10 CEUs Mean (SD)	≥ 11 CEUs Mean (SD)	t (df)	d
Adherence to standards	4.59 (.56)	4.84 (.38)	3.05 (120)**	.56
Accepted by practitioners	4.59 (.56)	4.84 (.38)	3.05 (120)**	.56
Minor violations	3.35 (.99)	3.73 (.91)	2.16 (122)*	.39

* $p < .05$; ** $p < .005$

the degree of design sensitivity conventionally accepted within the social sciences.[20]

A one-way ANOVA using Tukey HSD for paired comparisons was calculated on psychologists' subjective weights for 32 variables, comparing participants whose primary state of employment followed the *Daubert* standard (n = 38), the *Frye* standards (n = 50), or a hybrid/other standard (n = 37). No significant differences were obtained. Retrospective analysis indicated statistical power of .99, more than sufficient for most purposes in the social sciences (Muller & Benignus, 1992).

On the whole, psychologists' subjective weights were very consistent irrespective of how they were divided based on differences in training, education, and experience. Only 3 of 192 comparisons were significantly different, and the effect size for each was moderate. Despite controls for Type I error, the possibility that one or more of these significant findings is spurious must be considered. The three items in question do not share a strong theoretical relationship, and the important finding is the lack of difference across comparisons. In general, participants, regardless of training, education, and experience, are remarkably consistent in the subjective weights that they place on these variables when evaluating the scientific merit of testimony.

Research Goal 6

Determine what differences, if any, can be identified between psychologists with different training and different experience with the legal system in their estimates of the weights judges assign to these factors.

[20] Although conventional, some authors recommend a higher standard of statistical power, as high as .90 for some purposes (Muller & Benignus, 1992), whereas others consider retrospective power analysis to be largely uninformative and inappropriate for most purposes (Hoenig & Heisey, 2001).

Participants were grouped using the same six variables (and rationales) described under Research Goal 5, above. Independent t-tests (Bonferroni corrected) were calculated on psychologists subjective weights for 12 variables, comparing participants who have completed one or more full semester courses in psychology in law (n = 51) and those who have not completed any full semester courses in psychology and law (n = 73); comparing participants who have completed 10 or fewer continuing education seminars or workshops on the evaluation of scientific evidence for courtroom use (n = 52) and those who have completed 11 or more (n = 72); comparing participants who spend 50% or more of their professional time on court-related forensic activities (n = 55) and those who spend less than 50% of their time on such activities (n = 69); comparing participants who have testified in 15 or more criminal cases (n = 61) and those who have testified in fewer than 15 (n = 63); comparing participants who have testified in 15 or more civil cases (n = 61) and those who have testified in fewer than 15 (n = 64). Out of all of these comparisons, only significant two significant differences were obtained:

First, participants who had completed more continuing education on the evaluation of scientific evidence for the courtroom estimated that judges place more weight on whether the testimony is based on sufficient facts or data (mean = 4.62, SD = .60) than those who had completed less continuing education on the subject (mean = 4.15, SD = .91), t (121) = 3.40, $p < 01$, d = .62. Second, participants who had testified in more civil cases estimated that judges place more weight on general acceptance (mean = 4.66, SD = .60) than those who had testified less often in civil court (mean = 4.31, SD = .84), t (97) = 2.54, $p < 05$, d = .52.

A one-way ANOVA using Tukey HSD for paired comparisons was calculated on participants' estimated weights for judges on 12 variables, comparing participants whose primary state of employment followed the *Daubert* standard (n = 38), the Frye standards (n = 50), or a hybrid/other standard (n = 37). No significant differences were obtained.

On the whole, psychologists' estimated weights for judges were very consistent irrespective of how they were divided based on differences in training, education, and experience. Only 2 of 72 comparisons were significantly different, and the effect size for each was moderate. Despite controls for Type I error, the possibility that one or more of these significant findings is spurious must be considered. The two items in question do not share a strong theoretical relationship, and the overwhelming finding is the lack of difference across comparisons. In general, participants, regardless of training, education, and experience, are remarkably consistent in the estimated weights that they believe judges place on these variables when evaluating the scientific merit of testimony. As described under Research Goal 5, above, the design sensitivity for these comparisons was likely to adequate, or more than adequate, for the purposes of this dissertation.

Research Goal 7

Determine what differences, if any, can be identified between psychologists' self-reported weights and their estimates of weights assigned by judges to these factors.

Participants self-reported weights were compared to their estimated weights for judges by calculating dependent t-tests (Bonferroni corrected) on 12 variables. As can be seen in Table 10, participants reported that they place more weight on 11 of

the 12 variables than they believe judges do when evaluating the scientific merit of proposed testimony, with 9 comparisons reaching statistical significance. The notable exception was face validity, for which the reverse held. This suggests that psychologists believe that they give greater consideration to more pieces of information than do judges, particularly when it comes to information that may bear most directly on science. Retrospective analysis indicated statistical power of .99, more than sufficient for most purposes in the social sciences (Muller & Benignus, 1992).[21]

Research Goal 8

Examine the intercorrelation of participants' self-reported subjective weights.

The obtained sample size (n = 126) was not sufficient to complete an exploratory principle components analysis (PCA) with any confidence in the stability of the factor structure. A sample size of 200 can be sufficient to provide a stable component structure given moderate factor loadings (Velicer & Fava, 1998), but a sample size of 125 is lower than virtually all published guidelines for absolute minimum sample size (Guadagnoli & Velicer, 1988), nor does it meet the lowest published participant-to-variable ratios (5:1; Gorsuch, 1983). As an exploratory exercise, a PCA using Varimax (orthogonal) rotation was conducted, but the results can do little but serve the purpose of generating hypotheses for further research; any conclusions about the component structure derived from this analysis would be

[21] There is some dispute over the most appropriate way to estimate effect sizes for dependent t-tests (c.f., Dunlop, Cortina, Vaslow, & Burke, 1996; Rosnow & Rosenthal, 1996). The calculation of Cohen's d using the t value and pooled standard deviation is likely to over-estimate d, and possibly by a large margin, depending on the degree of correlation between scores (Dunlop, Cortina, Vaslow, & Burke, 1996). The calculation of d using the sample means and standard deviations is a more conservative estimate, and has been used here.

Table 10

Comparisons Between Participants' Self-Reported Weights and Participants' Estimated Weights for Judges.

Item[22]	Self-Reported Weights Mean (SD)	Estimated Weights for Judges Mean (SD)	t(df)	d
General acceptance	4.52 (.68)	4.50 (.72)	.272 (124)	
Error rate	4.15 (.78)	3.06 (1.02)	9.86 (123)*	1.20
Method can be tested	4.37 (.79)	3.35 (.73)	9.40 (123)*	1.34
Method has been tested	4.40 (.73)	3.63 (1.00)	7.34 (123)*	.77
Peer-reviewed publication	4.48 (.67)	3.85 (.94)	6.36 (123)*	1.15
Standards for administration	4.67 (.56)	3.73 (1.02)	9.58 (124)*	1.14
Due consideration to alternatives	4.51 (.66)	3.66 (.96)	8.96 (124)*	1.03
Based on Sufficient data	4.86 (.34)	4.35 (.82)	6.92 (124)*	0.81
Adherence to professional standards	4.68 (.64)	4.09 (1.01)	6.23 (123)*	1.12
Face validity	3.15 (1.03)	3.71 (1.14)	-4.70 (124)*	0.52
Parsimony	3.46 (.90)	3.36 (1.04)	1.10 (124)	
Construct Validity	4.16 (.73)	2.90 (1.02)	14.03 (123)*	1.42

* $P < .001$

[22] The full text of items has been presented in Tables 7, 8, and in Appendix B. They have been truncated here for ease of comparison.

tentative at best. One, two, and three components solutions were examined. The two-component solution was most interpretable. The rotated component solution is presented in Table 11.

Twenty of the thirty-two items loaded on a single component that accounted for 17.4% of the total variance. Six items loaded on a second component accounting for 16.6% of the total variance. Examination of the items loading on the two components suggests a single component formed almost entirely of items with a theoretical relationship to the trustworthiness of scientific evidence itself, and a second component formed of items suggesting deviation from basic standards of practice so severe that the competence or integrity of the expert could be questioned. Again, it should be emphasized that this analysis is highly exploratory in nature. Cross-validation on a larger sample would be required to draw meaningful conclusions, but the loading of most variables on a single component is consistent with psychologists' reports that they give moderate to great weight on virtually all factors under consideration.

Table 11

Exploratory Principle Components Analysis of Psychologists' Subjective Weights.

Item[23]	Rotated Component Loadings	
	Component 1	Component 2
General acceptance	.53	-.10
Error rate	.48	-.04
Method can be tested	.54	-.01
Method has been tested	.56	-.06
Published in peer-reviewed journals	.61	-.14
Standards for administration and interpretation	.54	.03
Due consideration for alternative explanations	.49	.18
Based on sufficient facts or data	.46	.14
Expert followed standards of practice	.46	.09
Method has face validity	.04	.33
Method has parsimony of explanation	.32	.39
Method has construct validity	.42	.10
Method is generally accepted	.53	-.10
Method is generally rejected	.48	-.04

[23] The full text of items has been presented in Table 7 and in Appendix B. They have been truncated here for ease of comparison.

Table 11, *cont.*

Exploratory Principle Components Analysis of Psychologists' Subjective Weights.

Item	Rotated Component Loadings	
	Component 1	Component 2
Method accepted by most but far from all	.54	-.01
Method rejected by most but far from all	.56	-.06
Method neither accepted nor rejected	.61	-.14
Practitioners accept; studies negative	.54	.03
Practitioners reject; studies positive	.47	.18
Error rate cannot be determined	.46	.14
Error rate has not been determined	.46	.09
Error rate is 20%	.04	.33
Error rate is 30%	.32	.39
Error rate is 50%	.42	.10
Error rate is 70%	.01	.83
Error rate is 80%	-.12	.82
Method has no standards for administration	-.02	.79
Expert adhered to standards for administration	.39	.39
Minor violations of standard procedures	-.03	.62
Major violations of standard procedures	-.06	.87
Gross and total violations of standard procedures	-.14	.84
Expert did not retain raw data underlying opinions	-.06	.78

DISCUSSION

Error Rate

Several questions on the maximum error rate for methods, individually and in combination, that participants would consider acceptable for use at trial were included for the purpose of collecting baseline data, but no hypotheses about participants' responses were formulated in advance. Although not identified as one of the primary research goals of the study, perhaps the most powerful and surprising findings of the dissertation are related to the issue of what level of error is acceptable for individual methods, or for methods used in combination, to be considered sufficiently trustworthy for admissibility.

First and foremost, a non-trivial percentage of participants endorsed maximum acceptable error rates for individual methods that are lower than the known error rates of many, if not most, procedures available in psychology today. For example, 44% of participants endorsed a maximum error rate of $\leq 10\%$ for a method to be introduced in civil court, and a full 56% of participants endorsed this standard in the criminal court setting. And while an error rate of 10% would be a very high standard for diagnostic or predictive purposes for any one test, it is not possible to determine how much more accurate participants believes individual tests should be! In designing the survey, and based on knowledge of the error rates typical within psychological testing, the author did not think it necessary or appropriate to provide response options that would allow participants to specify error rates lower than 10%. Consequently, we do not know whether these participants view an error rate of 1 in 100, 1 in 10,000, or zero as the maximum rate acceptable for the results of a method to be considered sufficiently

trustworthy for admission into evidence.[24] This is a stunning result, given that information on the accuracy of individual tests is generally published and readily available. This finding also raises the question of why psychologists would employ multiple measures for diagnostic or predictive purposes in the legal setting.

It would be fascinating, and potentially of great value, to conduct a follow-up survey within the same population, asking forensic psychologists to estimate the error rates of various commonly used methods (including those that they employ), and to compare those estimates with the error rates of those methods that have been established through research. If the results of the current dissertation are valid, then it is likely that forensic psychologists over-estimate, and perhaps to a large degree, the accuracy of the individual measures that they use.

The second interesting finding that resulted from the questions on error rates is that most participants endorsed a standard for the combined error rate of a group of procedures that equaled or exceeded the acceptable error rate for individual procedures. This raises questions as to how participants conceptualize the error rate of procedures and how they believe that the information obtained from different procedures is combined.

It is arguably a rational strategy to combine information from multiple sources in an effort to increase the accuracy of judgments, depending, of course, on how that information is combined. For example, a basic psychometric principle of test construction is that a combination of items, in a statistically derived scale, will be more reliable (i.e., have greater precision in measurement) than the individual items of

[24] In fact, one participant wrote a marginal note stating that he or she prefers tests to be *at least* 99% accurate in either a civil or criminal venue.

which it is comprised. As participants endorse cumulative error rates that are equal to, or less than, that of individual procedures taken separately, they may hold the view that psychologists are able to combine information in a way that allows for greater accuracy. For example, the belief may be that psychologists are able to integrate data in a way that takes advantage of the incremental validity[25] that some procedures may possess when combined with other tests within a battery of tests.

To the extent that psychologists believe that the error rate of a group of tests will be generally be equal to, or lower than, the error rate for the individual tests when the data is combined using subjective judgment, this belief is not supported by research; experts *may* gain some accuracy through the subjective combination of valid tests, but they are not able to do so optimally, and the effort to combine information in this manner may degrade accuracy even if all indicators have some degree of validity (Dawes, Faust, & Meehl, 1989; Faust, 2003; Grove & Meehl, 1996). It is also possible that the consideration of all of the data is a logical strategy if a disjunctive strategy is employed. For example, if standing any one variable is sufficient to deem testimony sufficiently trustworthy to admit into evidence, or sufficiently untrustworthy for admission, then all information could be considered in a non-configural manner that does not require complex data integration. Likewise, the utilization of a decision algorithm (e.g., a decision tree) could allow for a rational means of data integration

[25] Incremental validity refers to the degree that the addition of new information provides a statistically significant improvement in predictive or diagnostic accuracy. For example, one could attempt to combine the results of one of the best-validated intelligence tests (e.g., the Wechsler Adult Intelligence Scale) with a second test of intelligence for some purpose. Even if the second test is valid, the addition of the second test may not provide any improvement over the WAIS as a stand-alone measure. Incremental validity is maximized when a group of tests are each highly correlated with a criterion variable, but relatively orthogonal with one another. That is, each contributes valid information that is not redundant with the information obtained from other tests.

that would allow for all variables to be given at least some weight without resort to complex subjective data integration strategies.

The issue of cumulative error rate within a battery of tests taken together is an important one, given that tests are not typically used in isolation, and that a known error rate is one of the primary evaluative tools identified in the Daubert decision. To the extent that psychologists employ flexible batteries that have not been validated *as a combination*, the error rate of the battery is not known. The combination of tests could result in an increase or decrease in the error rate depending on how the information is combined. Flexible batteries that lack adequate standardization and validation have been successfully challenged in court due to the lack of supportive scientific evidence for the combination of tests (Reed, 1996). A recent case, *Baxter v. Temple* (2008), ruled in favor of the admissibility of unvalidated combinations of tests provided the individual components had been validated separately. However, this decision runs counter to basic psychometric theory and to the professional standards of the field: the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The *Standards*, which is an official APA policy document, specify that when a battery of tests is used for diagnostic purposes, the battery must be validated *as a whole*, irrespective of any validation research that may have been conducted on the individual components. It will take some time before the consequences of this unfortunate precedent unfold within the legal system.

A second possibility is that participants' high standards for the accuracy of methods in combination does not have any bearing on how participants believe psychologists *do* combine data from multiple sources, but reflects, instead, the standards that they believe *should* apply. In this case, participants may not believe that expert clinical judgment allows an individual to combine a variety of probabilistic indicators of varying quality in a way that improves judgmental accuracy, but believe that the addition of data is only warranted to the degree that accuracy is not degraded as a result. In this instance, one would expect a high rate of endorsement of actuarial methods amongst participants, but this was not directly assessed.

The final issue raised by the questions on error rates is how participants interpreted the questions themselves. The language of the items asked participants to specify the maximum error rate that they would consider acceptable for use at trial (*expressed as a percentage of errors*)? The wording is suggestive of sensitivity, specificity, and likelihood ratios, but could be taken in other ways, for example in relation to confidence intervals or magnitude of error. In addition, it is possible that some participants may have confounded the concept of error rate as it applies to the accuracy of tests or procedures with the concept of Type I and Type II error rates found in tests of statistical significance in research. The interpretation of "error rate" as the likelihood that a research finding is a statistical artifact and unlikely to be replicated is quite different from conceptualizing error rate as the number of correct, versus incorrect, decisions.[26]

[26] Setting this important conceptual difference aside, there is also the practical issue that a test may show a statistically significant relationship with some outcome and yet still achieve a level of accuracy that barely exceeds chance.

The acceptable error rate for a procedure could be evaluated in different ways. Likewise, what constitutes an "acceptable" error rate depends on the nature of the task involved. A high false positive rate could be acceptable in one context but not in another. Two participants articulated this conceptual difficulty in marginal comments. A refined and expanded study of error rate, taking these factors into account, is warranted.

Participants refused the error rate questions at a higher rate than any other group of items (11 – 13.5% of subjects failed to provide responses to these). Some participants wrote marginal notes stating that they did not understand the questions, and several argued at length that clinical assessment procedures do not have error rates, so that such questions are meaningless.

General Acceptance and a Sizeable Minority

As described above, participants' responses spanned the full range of response options. Participants offered a minimum threshold for a method to be "generally accepted" as low as 10% or less, and roughly 20% of the participants set the threshold at 40% or less. Likewise, participants offered a minimum threshold for a method to be accepted by a "sizeable minority" as high as 90% or more, and roughly 34% set the minimum threshold at greater than 50%. Conceptually, it is difficult to understand how a method can be considered "generally accepted" when it is endorsed by a minority within the field, or how a "sizeable minority" can be defined as exceeding 50% of the field.

Also notable is the frequency of extreme values endorsed by some participants: 10% of participants set the threshold for general acceptance at 90% of the field, a

standard unlikely to be met by most theories and methods within the field. In contrast, nearly 4% of participants endorsed 10% or less as the threshold for a sizeable minority, a standard that could be met by a wide range of theories and methods irrespective of their underlying scientific merit.

It is not clear how these results should be interpreted. It is possible that some participants were careless in their responses, which could cast doubt on the survey results as a whole. However, it is also possible that many respondents interpreted the questions accurately. In the latter case, forensic psychologists' understanding of what constitutes a majority consensus, or respectable minority viewpoint, on matters of scientific merit warrants a further examination. It is possible that some of the range in responses lies in differences in participants' interpretation of "general acceptance." For example, a specialized method in psychology might be unknown to 90% of the broader community of psychologists but accepted by 90% of the remaining 10% who have expertise in the relevant area.

Taking the median and modal responses for the two questions, the non-unanimous consensus for a method to be generally accepted would be endorsement by approximately 2/3 of the relevant field, and 1/3 would constitute a sizeable minority. However, the range of responses, and logical conundrums they produce, suggests that this inference cannot be taken as more than a hypothesis for further investigation with more refined methods.

Research Goal 1

Determine the self-reported weights assigned by psychologists to factors of potential relevance to evaluating the scientific trustworthiness of expert witness

testimony in the legal setting, i.e., to determine psychologists' subjective estimates of cue utilization when appraising scientific merit for courtroom use.

Participants indicated that they placed moderate to great weight on all variables, with the exception of face validity (which was still accorded some weight: a median of 3 on the five-point scale). This is an interesting finding given that it is extremely unlikely that these variables are all equally valid predictors of the scientific merit of proposed testimony. For example, whether a procedure is generally accepted probably has less direct bearing on the ultimate issue of trustworthiness than does an established, and acceptable, error rate. One wonders how many additional items could have been included in the survey with the same result, or what items would be appraised as carrying little or no weight in appraising scientific merit?

As discussed above in the context of cumulative error rate, above, the consideration of all available data, or even a limited pool of information of varying quality, does not necessarily increase judgmental accuracy. To the extent that a holistic subjective judgment is employed, accuracy is likely to degrade as additional information of relatively weaker predictive value is added. The use of actuarial methods eliminates many of the error sources of error that contribute to this phenomenon. For example, a properly developed actuarial method does not add new variables to existing ones unless inclusion of the new information yields gains in incremental validity.

Research on clinical judgment suggests that the holistic, subjective integration of data is the strategy most often employed (Faust & Ackley, 1998; Grove & Meehl, 1996), but that the consistent weighting of all variables as important is not a rational

strategy, nor is it consistent with how experts actually engage in decision-making. Given that decision-makers often have limited insight into how they evaluate and integrate information, it is unlikely that psychologists do, in fact, give equal, or nearly equal, consideration to all of the variables rated in this study (Aspel, Willis & Faust, 1998; Fisch, Hammond, Joyce & O'Reilly, 1981; Gauron & Dickenson, 1966; Kirwan, Chaput de Saintonge, Joyce & Currey; 1983; Nisbett & Wilson, 1977). Rather, studies suggest that decisions often rest in large part or entirely on a few variables, or even a single variable, despite subjective impressions to the contrary.

Participants' interpretation of the task might also be considered in evaluating self-reported weights. The questions specifically asked participants to indicate the weight that they would place on each variable, without reference to whether standing on the variable would support or undermine the trustworthiness of the proposed testimony. Review of survey responses and written comments in the margins suggested that some participants may have misinterpreted the task and responded to the items as if the Likert scale of "1 = no weight" to "5 = great weight" should be interpreted as a rating of the degree to which the variable would support admissibility. That is, some participants may have responded as if the scale ran from "1 = strongly supports exclusions" to "5 = strongly supports admission." As an example, it seems unlikely, and irrational, that an error rate of 80% would be given no weight when considering the scientific merit of a method, yet 35% of participants gave that response. Modification of the instructions for this portion of the survey, or survey items that explicitly include both dimensions (weight and direction) would help to ensure that weight and direction are not confused in subsequent studies.

It is worth noting that many of the items presented to participants for weighting were closely related, both conceptually and grammatically. One could question whether the relatively homogeneous weighting of these factors was influenced, at least in part, by limited variability in the item pool. Given that a large body of empirical research suggests that practitioners claim to consider all of the available data when making judgments (Faust, 2003; Faust & Ackley, 1998; Grove & Meehl, 1996) it seems unlikely that more variability in the scale items would have produced substantially greater variance in participants' weights. This is, of course, an empirical question that should be studied in the future.

In future research, an investigator might be challenged to find some variable that respondents do not deem important. In truth, this concern is not merely humorous or academic because optimizing decision accuracy is as much (if not more) a matter of rejecting or disregarding less valuable information as it is considering and incorporating valuable information. An insufficiently discerning view that everything is equally (or nearly equally) important is almost a surefire way to reduce the efficiency of decision making (Faust, 2003).

Research Goal 2

Evaluate the degree to which psychologists agree (or do not) as a group on their self-reported weights assigned to these factors of interest.

As described above, participants were in striking agreement in their impression of the subjective weights that they place on the variables they were asked to consider ($\alpha = .84$). This is not surprising after examination of participants' self-reported

weights. Statistical estimates of absolute agreement are necessarily high when there is so little variance in weights.

Research Goal 3

Determine the impressions psychologists have of the weights judges would assign to a subset of these to factors, i.e., to determine psychologists' estimates of the cue utilization of judges when appraising scientific merit for courtroom use.

Participants estimated that judges place some weight to great weight on all variables. As discussed under Research Goal 1, above, it is extremely unlikely that all variables are equally valid predictors of scientific merit, but the interpretation of these results is somewhat more ambiguous. Discussion of this finding will be more informative if integrated into the discussion of Research Goal 7, below, and will be addressed in that section. The same methodological concerns that applied to Research Goal 1 applied here as well.

Research Goal 4

Evaluate the degree to which psychologists agree (or do not) as a group on their estimates of the weights judges would assign to these factors.

The absolute agreement of weights assigned by participants to variables was assessed by calculating Cronbach's alpha coefficient ($\alpha = .84$). Participants were in agreement in their estimates of the weights that judges place on these variables when considering the scientific merit of proposed testimony.

Research Goals 5 and 6

Determine what differences, if any, can be identified between psychologists with different training and different experience with the legal system in the self-reported weights assigned to these factors.

Determine what differences, if any, can be identified between psychologists with different training and different experience with the legal system in their estimates of the weights judges assign to these factors.

The overwhelming result of comparisons across respondents is that, irrespective of how the obtained sample was divided, virtually no significant differences were identified despite adequate to excellent statistical power. The 5 differences between groups that achieved statistical significance out of 264 tests do not warrant interpretation. The meaningful result is that, taken as a whole, differences in training, education, and experience as measured by the survey did not have a meaningful effect on participants' self-reported weights or their estimates of judges' weights.

Research Goal 7

Determine what differences, if any, can be identified between psychologists' self-reported weights and their estimates of weights assigned by judges to these factors.

Based upon the comparison of participant's self reported weights to participants' estimated weights for judges, participants believed the place more weight on a broader range of variables than judges do when evaluating the scientific merit of proposed testimony, with 9 comparisons reaching statistical significance. The notable exception was face validity, for which the reverse held. This perception may have

some merit, as the role of the expert witness is, presumably, to assist the trier of fact (ie., the judge or jury deciding a case) in understanding scientific or technical material that they would not be able to otherwise. It could well be the case that forensic psychologists give greater consideration to a wider variety of variables relative to judges based on differences in scientific expertise.

The largest challenge to the validity of this interpretation is the finding that participants, by self-report, placed moderate to significant weight on virtually all variables surveyed and might well report that they place similar weight on a many others if given the opportunity to do so. A more discriminating eye toward the variables that are more valid predictors of scientific status is inconsistent with the view that most, if not all, variables are important.

An alternative explanation for participants' perception of differences in weights is one of attributional bias. This could be a non-motivated bias resulting from direct awareness experts have of their thought processes and the efforts they may engage in when trying to determine the trustworthiness of scientific evidence; the processes and efforts of the judge are inferred indirectly. An attributional bias could also be motivated, for example, participants could believe that they rate important evidence as more important than others do. In either case, judges, if asked to provide their own self-reported weights, and estimated weights for forensic psychologists, might perceive that they are the ones who give more weight to a wider range of information.

It is notable that, according to participants, judges place the greatest weight on the *Frye* standard (general acceptance), whether the testimony appears to be based on

sufficient data, and whether the expert has adhered to the professional standards of his or her field when evaluating the admissibility of expert witness testimony. These findings are consistent with those of Dahir, et al. (2005), who found that judges rated the qualifications of the expert, general acceptance, a general appraisal of trustworthiness and the credibility of the expert as the variables they most often considered when determining the admissibility of testimony. Gatowski et al. (2001), found that although there was no agreement amongst judges on how to weigh the *Daubert* factors, general acceptance was favored, albeit slightly.

The estimations of the weights that judges place on variables related to scientific merit provided by participants in this study and the convergent evidence from judges' ratings in other studies, raises an interesting issue with substantial implications for expert testimony and its role in the legal system. If, in fact, the admissibility of expert witness testimony is principally decided based on general acceptance and issues related to the credibility of experts, then judges may not be performing the gate-keeping function specified in the Federal Rules of Evidence and by states that have adopted the *Daubert* standard. This would suggest that, legislation and precedent to the contrary, *Frye* remains the dominant standard for determining the admissibility of expert witness testimony. If this is true, then the role of precedent in legal decisions may require serious reconsideration, the training of attorneys and expert witnesses modified, and the tactics of challenging the admissibility of opposing experts reconsidered.

Furthermore, if it is true that judges, due to the nature of their training and education, tend to make decisions based on general acceptance and credibility issues,

over-value face validity, and undervalue variables more directly related to scientific merit, then psychologists may be in a quandary when considering how to testify in court. If judges do, in fact, consistently make admissibility decisions based on less relevant or valid information, then how does one ethically negotiate the demand present testimony in a way that is both accurate and, at the same time, helpful to the court?

Research Goal 8

Examine the intercorrelation of participants' self-reported subjective weights

The obtained sample was insufficient to do more than identify a tentative hypothesis that participants distinguish between issues related to scientific merit and issues related to gross professional misconduct or negligence. This is mildly interesting in that, as noted above, prior research suggests that judges may evaluate admissibility based on an appraisal of trustworthiness, dominated by the *Frye* standard, and expert credibility. Were this highly exploratory finding to be replicated on a larger sample, an interesting possibility emerges: that the structure of the appraisal of admissibility is two-dimensional for both judges and forensic psychologists. Each group may make those determinations based on different factors (for example, a judge may value general consensus and face validity whereas an expert may value error rate and construct validity for assessing the merit dimension), but there decision process is identical; the information and the source become the two criteria under consideration.

CONCLUSIONS

As an exploratory effort, the dissertation has met all of the primary and secondary research goals outlined, and demonstrates that research on the cue utilization of forensic psychologists evaluating the scientific merit of research is both feasible and promising. The strength of the conclusions that can be drawn from the results must be tempered due to the modest return rate, and the possibility that unintended ambiguity led to spurious variance in participants' interpretation of several survey items. However, the results are informative, and suggest a variety of avenues for follow-up research. The clarification of the elements that forensic actors utilize in weighing the scientific merit of proposed testimony, and how such decisions are made, can only be of benefit to society by promoting outcomes that are just and well-informed by science that rests on a solid foundation.

An understanding of the acceptable error rate for expert witness testimony is critical. At present there is little consensus in the courts or among expert witnesses on the conceptualization of error rates, or what level of error passes some minimum threshold for admissibility. In the current study, a sizeable majority of participants endorsed an error rate of 20 % or less as an appropriate standard for admissibility. Few participants endorsed an error rate of greater than 30 %. This would suggest that, depending on the goal of testing, forensic psychologists believe that the sensitivity, specificity, or predictive power (whichever statistic is most appropriate in the given context) for tests should be at least .80, and possibly higher. Beyond this, the large number of participants who endorse error rates lower, or potentially far lower, than what the procedures within the social sciences can typically offer is problematic on a

variety of levels. Direct comparisons of the accuracy rates endorsed by forensic psychologists, the accuracy rates they believe they achieve with their procedures of choice, and the empirically established accuracy rates of those procedures is perhaps the most urgent next step for additional study.

When considering a group of tests in combination, the finding that forensic psychologists endorse a maximum error rate that is roughly the same as that endorsed for measures taken in isolation raises an important issue with regard to the need for standardized test batteries that are interpreted using actuarial methods. Given the limitations on human judgment and basic principles of psychometrics, it is unlikely that flexible batteries of tests, interpreted through subjective judgment, can produce error rates that are lower, or substantially lower, than those of the best component measures. The effort to combine data holistically using subjective judgment is, in fact, more likely to reduce judgmental accuracy rather than improve it.

Research aimed at the conceptualization of error rates is clearly warranted. Standards of error as defined by sensitivity, specificity, predictive power, standard error of measurement, standard error of estimate, and predictive accuracy should all be examined to determine the limits of acceptable error when error is expressed in different ways. In addition, it would be valuable to compare the threshold for acceptable error rates as perceived by forensic psychologists, experts in other fields, attorneys, and judges. It is likely that different actors in forensic contexts will endorse different standards, and the standards may also vary depending on the field in question. For example, during preliminary interviews, one judge initially endorsed a relatively high threshold for admissibility. When told that this standard probably

exceeded that typical in the social sciences, she reconsidered, stating that if that were the case, then perhaps a lower standard would be appropriate when considering the admissibility of testimony in these fields.

A second avenue for additional research is examination of the concepts of "general acceptance," or the *Frye* standard, and of a sizeable, substantial, or respected "minority." Participant's responses suggest, however crudely, that "general" acceptance is attained when a method or theory is accepted by approximately 70% of the relevant scientific community, and that a "sizeable" minority is constituted by approximately 30% of the field. However, the wide range of responses to these questions, and the number of responses that defy logical interpretation, indicate that these results may not be valid. Research aimed at clarifying these concepts, and comparing the views of forensic psychologists to those of judges, attorneys, and experts from other fields is indicated.

A striking finding from the study is the degree to which forensic psychologists are consistent in their reports of subjective cue utilization, and in their belief that they give significant consideration to all of the variables included in the survey. This reveals what may be a substantial flaw in how forensic psychologists weigh and integrate information, or at least in how they perceive they weigh and integrate information. It is a virtual certainty that the variables surveyed in this study are cannot be equally valid predictors of true scientific status with roughly equivalent utility. As discussed at length above, this belief is inconsistent with research on human judgment and cue utilization, which demonstrates that such judgments are usually made using a small subset of the cues available, irrespective of subjective reports. To the extent that

forensic psychologists do, in fact, believe that they integrate "all of the data," including the variables surveyed and many others that could have been included in the study but were not, this approach requires close examination for both pragmatic and ethical reasons.

The success in gaining information on subjective weights placed on variables of potential relevance to the admissibility of scientific evidence is exciting, for it provides evidence for the feasibility of a program of research on the cue utilization of forensic psychologists, attorneys, and judges. In refining our understanding of cue utilization, future research should also clarify the distinction between the weight attached to variables, and whether standing on variables supports admissibility or rejection of the testimony. Additional research can be aimed at comparisons between the subjective and statistically-determined cue utilization within each population, and comparisons across these populations as well. To the extent that these efforts are, themselves, successful the theoretical possibility of developing boot-strapped statistical models of the admissibility decisions of judges may become a practical possibility.

Appendix A

Informed Consent Document

The University of Rhode Island
Department of Psychology
10 Chafee Rd., Suite 8, Kingston, RI 02881

Title of Project: Factors Influencing the Evaluation of Expert Witness Testimony in the Behavioral Sciences Under *Daubert*.

PLEASE RETAIN THIS PAGE FOR YOUR RECORDS.

You have been asked to take part in the research project described below.

The purpose of this study is to examine how psychologists evaluate the scientific merit of expert witness testimony for the courtroom. If you decide to take part in this study, your participation will involve completing a survey on factors potentially relevant to the evaluation of scientific merit in legal settings.

There will be no direct benefits or compensation provided to you, but you will be provided with a summary of the results at your request once the study is complete. Your participation will help increase our knowledge regarding the evaluation of expert witness testimony, as well as help me to fulfill the requirements for the Ph.D. in Clinical Psychology.

Participation in this study is anonymous and confidential. Your involvement and responses cannot be identified. Any published reports will be based on group data only. Participation in this study is voluntary. You may decline to participate, refuse to answer individual questions, or quit the survey at any time.

You may opt to provide an e-mail or postal address in order to receive a summary of the results once the study is complete using the separate postage-paid card enclosed. This contact information will not be recorded in the same database as survey responses, and cannot be linked to the responses of any individual participant.

This study has been reviewed and approved by the University's Institutional Review Board on Human Subjects. Participation in this study is not expected to be harmful or injurious to you.

If you have any concerns or questions at any time, you should write or call Kenneth Heard, M.A. at 401-374-0167 or David Faust, Ph.D. at 401-874-4237, the people mainly responsible for this study. You may also contact the University of Rhode Island's Vice Provost for Graduate Studies, Research and Outreach, 70 Lower College Road, Suite 2, URI, Kingston, RI, 401-874-4328 with any questions or concerns you may have.

You must be 18 years of age to participate in this study. The act of completing and returning the survey affirms that any questions have been answered to your satisfaction and that you consent to participate.

APPENDIX B

Survey

FACTORS INFLUENCING THE EVALUATION OF

EXPERT WITNESS TESTIMONY IN THE BEHAVIORAL SCIENCES

UNDER *DAUBERT*

Part 1:

1) Please indicate your academic degree(s): Ph.D. Psy.D Ed.D. J.D. Other (please describe): _____

2) Did you obtain your degree through a program specializing in psychology and the law or forensic psychology? Yes No

3) How many full-semester college or university courses in forensic psychology or psychology and law have you completed? 0 1 2 3 4 5 ›5

4) How many brief professional or continuing education workshops or seminars on evaluating the trustworthiness or merit of scientific evidence in the legal setting have you completed? 0 1 – 10 11 – 20 21 – 30 >30

5) Please find the state in which you are primarily employed and indicate which column it is listed:

Group A		
AK	AR	CO
CT	DE	HI
ID	IN	IA
KY	LA	ME
MS	MT	NE
NH	NM	OH
OK	OR	RI
SC	SD	TN
TX	UT	VT
WV	WY	

Group B		
AL	AZ	CA
DC	FL	IL
KS	MD	MI
MN	NJ	PN
	WA	

Group C		
GA	MA	MS
NV	NY	NC
ND	VA	WI

6) Approximately what percentage of your professional time is spent on forensic work directly related to courtroom activities (including, for example, preparation, assessments, report-writing, consultation and testimony)?

0% ≤10% 20% 30% 40% 50% 60% 70% 80% ≥90%

7) In the past ten years, in approximately how many *criminal* cases have you been offered as an expert witness?

0 1 – 5 6 – 10 11 – 15 ›15

a) In how many *criminal* cases has a motion to exclude your expert testimony been made based on issues of scientific trustworthiness or merit?

0 1 2 4 5 6 7 8 9 ≥10

b) How many of these motions (if any) resulted in the exclusion of your testimony?

0 1 2 4 5 6 7 8 9 ≥10

8) In the past ten years, in approximately how many *civil* cases have you been offered as an expert witness?

0 1 – 5 6 – 10 11 – 15 ›15

a) In how many *civil* cases has a motion to exclude your testimony been made based on issues of scientific trustworthiness or merit?

0 1 2 4 5 6 7 8 9 ≥10

b) How many of these motions (if any) resulted in the exclusion of your testimony?

0 1 2 4 5 6 7 8 9 ≥10

9) If you have provided expert witness testimony in court cases in the past ten years, please indicate approximately what percentage of these cases were in:

a) Federal Court

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

b) State Court

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

10) If an expert administers a group of tests or assessment procedures, what is the maximum individual error rate for any one method in that group that you would consider acceptable for use in a *civil* trial (expressed as a percentage of errors)?

≤10% 20% 30% 40% 50% 60% 70% 80% ≥90%

11) If an expert administers a group of tests or assessment procedures, what is the maximum individual error rate for any one method in that group that you would consider acceptable for use in a *criminal* trial (expressed as a percentage of errors)?

≤10% 20% 30% 40% 50% 60% 70% 80% ≥90%

12) If an expert administers a group of tests or assessment procedures, what is the maximum cumulative error rate for the group as a whole that you would consider acceptable for use in a *civil* trial (expressed as a percentage of errors)?

≤10% 20% 30% 40% 50% 60% 70% 80% ≥90%

13) If an expert administers a group of tests or assessment procedures, what is the maximum cumulative error rate for the group as a whole that you would consider acceptable for use in a *criminal* trial (expressed as a percentage of errors)?

≤10% 20% 30% 40% 50% 60% 70% 80% ≥90%

14) What percentage of agreement within a professional field surpasses the threshold for a method to be considered generally accepted in that field?

≤10% 20% 30% 40% 50% 60% 70% 80% ≥90%

15) What percentage of agreement within a professional field surpasses the threshold for a method to be considered accepted by a sizeable minority in that field?

≤10% 20% 30% 40% 50% 60% 70% 80% ≥90%

16) Please rate how familiar you are with the criteria for evaluating the trustworthiness or merit of scientific evidence described in *Daubert v. Merrell Dow Pharmaceuticals*?

Not Familiar Somewhat Familiar Moderately Familiar Very Familiar

Part 2: For the following questions, please assume that a psychologist has collected data (e.g., through chart review, psychological assessments, and clinical interviews) and then used this information to make interpretations in preparation for expert testimony in a legal case. For the purposes of this study, the psychologist's data collection and interpretation will be referred to as the psychologist's "*method.*"

In addition, please assume that you have been asked to participate in a neutral "blue ribbon" panel to advise the court on the scientific merit or trustworthiness of proposed expert witness testimony for courtroom use.

Please estimate how much weight **you** would place on each of the following factors when evaluating the scientific merit or trustworthiness of evidence for courtroom use. In your practice you may generally consider variables in combination and evaluate information in a holistic manner. I recognize that it may be difficult or somewhat artificial to consider each item in isolation, but please attempt to do so for the purposes of this study. Rate each item on the following scale:

1	2	3	4	5
No Weight.	Little Weight	Some Weight	Moderate Weight	Great Weight

1) Whether or not the method is generally accepted within the relevant scientific community.
2) The error rate of the method.
3) Whether or not the method can be tested.

1	2	3	4	5
No Weight.	Little Weight	Some Weight	Moderate Weight	Great Weight

4) Whether or not the method has been tested.
5) Studies of the method have been published in peer-reviewed sources.
6) The presence of, and conformity to, standards for the administration and interpretation of the method.
7) Whether or not the expert has given due consideration to viable alternative explanations for the results obtained by the method.
8) Whether or not the testimony appears to be based on sufficient facts or data.
9) Whether or not the expert has followed standards of practice adopted in his or her field.
10) The degree to which the method has face validity.
11) The degree to which the testimony employs parsimony of explanation.
12) The degree to which the method has construct validity.

Part 3: For the following questions, please assume that a psychologist has collected data (e.g., through chart review, psychological assessments, and clinical interviews) and then used this information to make interpretations in preparation for expert testimony in a legal case. For the purposes of this study, the psychologist's data collection and interpretation will be referred to as the psychologist's "*method.*"

In addition, please assume that you have been asked to participate in a neutral "blue ribbon" panel to advise the court on the scientific merit or trustworthiness of proposed expert witness testimony for courtroom use.

Please estimate how much weight you think **a judge** would place on each of the following factors when evaluating the scientific merit or trustworthiness of evidence for courtroom use. In your practice you may generally consider variables in combination and evaluate information in a holistic manner. I recognize that it may be difficult or somewhat artificial to consider each item in isolation, but please attempt to do so for the purposes of this study. Rate each item on the following scale:

1	2	3	4	5
No Weight.	Little Weight	Some Weight	Moderate Weight	Great Weight

13) Whether or not the method is generally accepted within the relevant scientific community.

14) The error rate of the method.

15) Whether or not the method can be tested.

1	2	3	4	5
No Weight.	Little Weight	Some Weight	Moderate Weight	Great Weight

16) Whether or not the method has been tested.

17) Studies of the method have been published in peer-reviewed sources.

18) The presence of, and conformity to, standards for the administration and interpretation of the method.

19) Whether or not the expert has given due consideration to viable alternative explanations for the results obtained by the method.

20) Whether or not the testimony appears to be based on sufficient facts or data.

21) Whether or not the expert has followed standards of practice adopted in his or her field.

22) The degree to which the method has face validity.

23) The degree to which the testimony employs parsimony of explanation.

24) The degree to which the method has construct validity.

Part 4: For the following questions, please assume that a psychologist has collected data (e.g., through chart review, psychological assessments, and clinical interviews) and then used this information to make interpretations in preparation for expert testimony in a legal case. For the purposes of this study, the psychologist's data collection and interpretation will be referred to as the psychologist's "*method.*"

In addition, please assume that you have been asked to participate in a neutral "blue ribbon" panel to advise the court on the scientific merit or trustworthiness of proposed expert witness testimony for courtroom use.

Please estimate how much weight **you** would place on each of the following factors when evaluating the scientific merit or trustworthiness of evidence for courtroom use. In your practice you may generally consider variables in combination and evaluate information in a holistic manner. I recognize that it may be difficult or somewhat artificial to consider each item in isolation, but please attempt to do so for the purposes of this study. Rate each item on the following scale:

1	2	3	4	5
No Weight.	Little Weight	Some Weight	Moderate Weight	Great Weight

25) The method is generally accepted in its field.
26) The method is generally not accepted by most members in its field.
27) The method is accepted by most, but far from all, members in its field.
28) The method is rejected by most, but far from all, members in its field.
29) The method is neither clearly accepted nor rejected in its field.

1	2	3	4	5
No Weight.	Little Weight	Some Weight	Moderate Weight	Great Weight

30) The method is generally accepted by practitioners (i.e., psychologists providing assessment and treatment) but research studies are generally negative.

31) The method is generally not accepted by practitioners (i.e., psychologists providing assessment and treatment) but research studies are generally positive.

32) There is no way for the method's error rate to be determined.

33) The error rate of the method could be determined but it has not been.

34) The method has an error rate of $\leq 20\%$.

35) The method has an error rate of 30%.

36) The method has an error rate of 50%.

37) The method has an error rate of 70%

38) The method has an error rate of $\geq 80\%$.

39) The method does not have standard administration procedures.

40) The expert has followed standard administration procedures for the method.

41) The expert has made minor violations of standard administration procedures for the method.

42) The expert has made significant violations of standard administration procedures for the method.

43) The expert has made total and gross violations of standard administration procedures for the method.

44) The expert did not retain the raw data underlying his or her opinions.

Thank you for taking the time and effort to participate in my dissertation research by completing and returning this survey. I greatly appreciate your help and look forward to sharing the results with you on completion of the study.

APPENDIX C

Cover Letter to Study Participants

Kenneth Heard, M.A.
Department of Psychology
University of Rhode Island
10 Chafee Rd., Ste 8
Kingston, RI 02881

[Date]

[Participant Name]
[Title]
[Address 1]
[Address 2]
[City, State, Zip Code]

Dear [Participant Name]:

I am writing to invite you to participate in a study I am completing towards the fulfillment of the requirements for my doctoral degree in clinical psychology at the University of Rhode Island. The goal of this study is to examine variables related to the evaluation of science for courtroom use.

You have been randomly selected from the membership of the American Psychology-Law Society through the APA Center for Psychology Workforce Analysis and Research as potential participant in this study due to your interest in psychology and the law. This study has been reviewed and approved by that office and by the University of Rhode Island's Institutional Review Board on Human Subjects.

If you choose to participate, you will be asked to complete a brief survey about the weight that you would place of factors related to the quality or merit of scientific evidence, and to give your impressions of the weight a judge might place on these factors.

The complete details of the study are described in the enclosed Informed Consent document and in the instructions for the survey. I will be more than happy to answer any questions you may have at any point during your involvement in this study.

I am unable to provide any direct compensation for your time and effort should you choose to participate, but I do hope that this study will make a meaningful contribution to our knowledge in this area, and you will have my gratitude for your help in completing my degree. I will be happy to provide you with the results of the study upon completion.

Thank you and I hope you will consider sharing your expertise.

Sincerely,

Kenneth Heard, M.A.

APPENDIX D

Text of Reminder Card

Dear [Participant],

Approximately two weeks ago, I mailed you a survey on the evaluation of science for the courtroom that I am undertaking as part of my doctoral dissertation at the University of Rhode Island. If you have had the opportunity to return this, I thank you for your time and consideration. If you desire to participate but have mislaid or did not receive the original mailing, please contact me at 401-374-0167, via e-mail at ekenheard@msn.com, or by mail at the Department of Psychology, University of Rhode Island, 10 Chafee Rd., Suite 8, Kingston, RI 02881, and I will mail a second copy to you right away.

Sincerely,

Kenneth Heard, M.A.

APPENDIX E

Text for Request for Results Card

Dear Mr. Heard,

Please forward a copy of the results of your dissertation on the evaluation of scientific evidence to my attention at the e-mail or postal address provided below.

Kenneth Heard
Department of Psychology
University of Rhode Island
10 Chafee Rd., Ste 8
Kingston, RI 02881

BIBLIOGRAPHY

Aspel, A. D., Willis, W. G., & Faust, D. (1998). School psychologists' diagnostic decision-making processes: Objective-subjective discrepancies. *Journal of School Psychology, 36,* 137 - 149.

Bernstein, D. E., & Jackson, J. D. (2004). The *Daubert* trilogy in the states. *Jurimetrics, 44*, 351 – 366.

Bradford-Hill, A. (1966). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58,* 295 – 300.

Brodsky, S. L. (1999). *The expert expert witness: More maxims and guidelines for testifying in court.* Washington, DC: American Psychological Association.

Buchdahl, G. (1970). History of science and criteria for choice. In: R. H. Stuewer (ed.) *Minnesota Studies in the philosophy of science. Vol. 5. Historical and philosophical perspectives of science.* (pp. 204 – 230). Minneapolis: University of Minnesota Press.

Bursoff, D. N. (1999). *Table of cases.* Unpublished document.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance, 27,* 411 – 422.

Ceci, S. J., & Hembrooke, H. (Eds.) (1998). *Expert witnesses in child abuse case: What can and should be said in court.* Washington, DC: American Psychological Association.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method.* New York: Harcourt, Brace & World

Dahir, V. B., Richardson, J. T., Ginsberg, G. P., Gatowski, S. I, Dobbin, S. A., &

Merlino, M.L. (2005). Judicial application of *Daubert* to psychological syndrome and profile evidence: A research note. *Psychology, Public Policy, and Law 11,* 62 – 82.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U. S. 579 (1993).

Dana, J., & Dawes, R. M. (In press). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics.*

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist, 34,* 571 – 582.

Dawes, R. M. (1986). Forecasting one's own preference. *International Journal of Forecasting, 2,* 5 – 14.

Dawes, R. M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95 – 106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668 –1674.

Dixon, L., & Gill, B. (2001). *Changes in the standards for admitting expert evidence in federal civil cases since the Daubert decision.* Santa Monica, CA: RAND Institute for Civil Justice.

Dixon, L., & Gill, B. (2002). Changes in the standards for admitting expert evidence in federal civil cases since the *Daubert* decision. *Psychology, Public Policy and Law, 8,* 251 - 308.

Dunlop, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of

experiments with matched-groups or repeated measures designs. *Psychological Methods, 1*, 170 – 177.

Dyer, F. J., & McCann, J. T. (2000). The Millon Clinical Inventories, research critical of their forensic application and *Daubert* criteria. *Law and Human Behavior, 24,* 487 -497.

Faust, D. (1984). *The limits of scientific reasoning.* Minneapolis, MN: University of Minnesota Press.

Faust, D. (1997). Of science, meta-science, and clinical practice: The generalization of a generalization to a particular. *Journal of Personality Assessment, 68,* 331 – 354.

Faust, D., & Ackley, M. A. (1998). Did you think it was going to be easy? Some methodological suggestions for the investigation and development of malingering-detection techniques. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 1 – 54). New York: Plenum

Faust, D. & Heard, K. V. (2003a). Objectifying subjective injury claims. In I. Schultz and D. Brady (Eds.). *Psychological Injuries at Trial [electronic resource]* (pp. 1686 - 1705). Chicago, IL: American Bar Association.

Faust, D. & Heard, K. V. (2003b). Biased experts: Some practical suggestions for identifying and demonstrating unfair practices. In I. Schultz and D. Brady (Eds.). *Psychological Injuries at Trial [electronic resource]* (pp. 1706 -1739). Chicago, IL: American Bar Association.

Faust, D., & Meehl, P. E. (1992). Using scientific methods to resolve questions in the

history and philosophy of science: Some illustrations. *Behavior Therapy, 23,* 195 – 211.

Faust, D., & Meehl, P. E. (2002). Using meta-scientific studies to clarify or resolve questions in the philosophy and history of science. *Philosophy of Science 69,* S185 – S196.

Federal Rules of Evidence (1974). Washington: USGPO

Federal Rules of Evidence (2000). Washington: USGPO

Feyerabend, P. (1993). *Against Method.* New York: Verso

Fisch, H. U., Hammond, K. R., Joyce, C. R. B., & O'Reilly, M. (1981). An experimental study of the clinical judgment of general physicians in evaluating and prescribing for depression. *British Journal of Psychiatry, 138,* 100 - 109.

Frye v. United States 292 F. 1013 (D.C. Cir. 1923).

Garb, H. N. (1999). Call for a moratorium on the use of the Rorschach Inkblot Test in clinical and forensic settings. *Assessment, 6,* 313 – 317.

Gatowski, S. I., Dobbins, S. A., Richardson, J. T., Ginsberg, G. P., Merlino, M. L., & Dahir, V. (2001). Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-*Daubert* world. *Law and Human Behavior, 25,* 433 – 458.

Gauron, E. F., & Dickinson, J. K. (1966). Diagnostic decision making in psychiatry. *Archives of General Psychiatry, 14,* 225-237.

General Electric Co. v. Joiner, 522 U. S. 136 (1997).

Goldberg, L. R. (1965). Diagnosticians versus diagnostic signs: The diagnosis of

psychosis versus neurosis from the MMPI. *Psychological Monographs, 79,* (9 whole No 602).

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgment. *American Psychologist, 23*, 483 – 496.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422 – 432.

Goldberg, L. R. (1976). Man versus model of man: Just how conflicting is the evidence? *Organizational Behavior and Human Performance, 16,* 13 – 22.

Goodman-Delahunty, J. (1997). Forensic psychological expertise in the wake of *Daubert*. *Law and Human Behavior, 21,* 121 – 140.

Gorman, B. J. (1999). Facilitated communication: Rejected in science, accepted in court – A case study and analysis of the use of FC evidence under *Frye* and *Daubert*. *Behavioral Sciences and the Law, 17,* 517 – 541.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Groscup, J. L., Penrod, S. D , Studebaker, C. A., Huss, M. T., &.O'Neil, K. M. (2000). The effects of *Daubert v. Merrell Dow Pharmaceuticals* on the admissibility of expert testimony in state and federal criminal cases. *Psychology, Public Policy & Law, 8,* 339-372.

Grove, W. M., & Barden, R. C. (1999). Protecting the integrity of the legal system: The admissibility of testimony from mental health experts under *Daubert/Kumho* analyses. *Psychology, Public Policy and Law, 5*, 224 – 242.

Grove, W. M., Barden, R. C., Garb, H. N., & Lilienfeld, S. O. (2002). Failure of

Rorschach-Comprehensive-System-based testimony to be admissible under the *Daubert-Joiner-Kumho* standard. *Psychology, Public Policy, & Law, 8,* 216 – 234.

Grove, W. M. & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293 – 323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B.E., & Nelson, N. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19 - 30.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. Psychological Bulletin, 103, 265-275

Gutheil, T. G., & Stein, M. D. (2000). *Daubert*-based gatekeeping and psychiatric/psychological testimony in court: Review and proposal. *Journal of Psychiatry and Law, 28,* 235 – 251.

Heard, K. V., & Faust, D. (2003). Expert witness, psychological aspects. In N. Smelser & P. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences.* London: Elsevier.

Heath, I. (2000). Does expert witness testimony regarding the child sexual ause accommodation syndrome (CSAAS) meet the *Daubert* standard for admissibility of expert testimony? *Journal of the American Academy of Psychiatry and the Law, 28,* 248 – 249.

Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, N.J.: Prentice-Hall.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*, 19 – 24.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57*, 116 – 131.

Jenkins, J. A. (2002). *Quality of sentencing in the Rhode Island courts.* Rhode Island Supreme Court, Committee on Women and Minorities in the Courts. Providence, RI: Author.

Johnson, M. T, Krafka, C., & Cecil, J. S. (2000). *Expert testimony in federal civil trials: A preliminary analysis*. Washington, DC: Federal Judicial Center.

Kirwan, J., Chaput de Saintonge, D., Joyce, C., & Currey, H. (1983). Clinical judgment in rheumatoid arthritis. II. Judging "current disease activity" in clinical practice. *Annals of the Rheumatic Diseases, 42,* 648 - 651.

Krafka, C., Dunn, M.A., Johnson, M.T., Cecil, J.S., & Miletich, D., (2002). Judge and attorney experiences, practices, and concerns regarding expert testimony in federal civil trials. *Psychology, Public Policy, & Law, 8*, 309 - 332.

Kuhn, T. (1968). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kumho Tire Co. v. Patrick Carmichael, 119 S. Ct. 1167 (1999).

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27 – 66.

Lally, S. J. (2001). Should human figure drawings be admitted into court? *Journal of Personality Assessment, 76,* 135 – 149.

Lipton, J. P. (1999). The use and acceptance of social science evidence in business litigation after *Daubert. Psychology, Public Policy, and Law, 5,* 59 – 77.

Margenau, H. (1950). *The nature of physical reality.* New York: McGraw Hill.

McCann, J. T. (1998). Defending the Rorschach in court: An analysis of admissibility using legal and professional standards. *Journal of Personality Assessment, 70,* 125 – 144.

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1971/1991). Law and the fireside inductions: Some reflections of a clinical psychologist. In C. A. Anderson and K. Gunderson (Eds.) *Paul E. Meehl: Selected Philosophical and Methodological Papers* (pp. 440 – 480). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1973). Why I do not attend case conferences. In P. E. Meehl, *Psychodiagnosis: Selected Papers* (pp. 225 – 302). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1978/1991). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. In C. A. Anderson and K. Gunderson (Eds.) *Paul E. Meehl: Selected Philosophical and Methodological Papers* (pp. 1 – 42). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370 – 375.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195 – 244.

Meehl, P. E. (1992). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports, 71*, 339 – 467.

Meehl, P. E. (1997). The problem is epistemology, not statistics. In L. Harlow & S. Muliak (Eds.). *What if there were no significance tests?* New York: Lawrence Erlbaum.

Meehl, P.E. (2002). Cliometric metatheory: II. Criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports, 91 (Monograph supplement 1-V91),* 339 – 404.

Meehl, P.E. (2004). Cliometric metatheory III: Peircean consensus, verisimilitude, and asymptotic method. *British Journal for the Philosophy of Science, 55,* 615-643.

Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (1997). *Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers (2nd ed.).* New York: Guilford.

Meltzoff, J. (1998). *Critical thinking about research: Psychology and related fields*. Washington, D.C.: American Psychological Association.

Muller, K. E., & Benignus, V. A. (1992). Increasing scientific power with statistical power. *Neurotoxicology and Teratology, 14*, 211 – 219.

National Academy of Sciences (2000). *Scientific evidence workshop: Science, technology, and law program.* Transcript of workshop, (9/17/2000). Washington, D.C.: Author.

Newton-Smith, W. H. (1981). *The rationality of science.* New York: Routledge.

Nisbett, R., & Wilson, T. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Bulletin, 84,* 231 - 259.

Nordberg, P.B. (2006). *Psychologists & Psychiatrists.* Retrieved February 21, 2006, from Daubert on the Web: http://www.daubertontheweb.com

O'Connor, M. & Krauss, D. (2001, Winter) Legal update: New developments in Rule 702. *American Psychology-Law Society News ,21,(1),* 1 – 4, 18.

Park, R. L. (2003). The seven warning signs of bogus science. *Chronicle Review, 49 (21)*, B20.

Penrod, S. D., Fulero, S. M. & Cutler, B. L. (1995). Expert psychological testimony in the United States: A new playing field? *European Journal of Psychological Assessment, 11,* 65 – 72.

Pope, K. S. (1998). Pseudoscience, cross-examination, and scientific evidence in the recovered memory controversy. *Psychology, Public Policy, and Law, 4,* 1160 – 1181.

Popper, K. (1959). *The Logic of Scientific Discovery.* New York: Hutchinson.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York: Routledge.

Reed, J. E. (1996). Fixed vs. flexible neuropsychological test batteries under the *Daubert* standard for the admissibility of scientific evidence. *Behavioral Sciences and the Law, 14,* 315 – 322.

Rescher, N. (1990). *Aesthetic factors in natural science.* Lanham, MD: University Press of America.

Ritzler, B., Erard, R., & Pettigrew, G. (2002). Protecting the integrity of Rorschach expert witnesses: A reply to Grove and Barden (1999) re: The admissibility of testimony under *Daubert/Kumho* analysis. *Psychology, Public Policy & Law, 8,* 201 – 215.

Rogers, R., Salekin, R. T., & Sewell, K. W. (1999). Validation of the Millon Clinical Multiaxial Inventory for Axis II disorders: Does it meet the *Daubert* standard? *Law & Human Behavior, 23,* 425 – 443.

Rosenthal, R., Hiller, J. B., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (2001). Meta-analytic methods, the Rorschach, and the MMPI. *Psychological Assessment, 13,* 449 – 451.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331-340.

Rotgers, F., & Barrett, D. (1996). *Daubert v. Merrell Dow* and expert testimony by clinical psychologists: Implications and recommendations for practice. *Professional Psychology: Research & Practice, 27,* 467 – 474.

Ruscio, J. (2006). *Critical thinking in psychology: Separating sense from nonsense.* Belmont, CA: Thompson – Wadsworth.

Sagan, C. (1996) *The demon-haunted world: Science as a candle in the dark.* New York: Random House.

Salmon, W. (1998). *Causality and scientific explanation.* New York: Oxford.

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual prediction methods. *American Journal of Sociology, 48,* 593 – 602.

Saxe, L., & Ben-Shakhar, G. (1999). Admissibility of polygraph tests: The application of scientific standards post-*Daubert*. *Psychology, Public Policy, and Law, 5,* 203 – 223.

Schaffner, K. F. (1970). Outlines of a logic of comparative theory evaluation with special attention to pre- and post-relativistic electrodynamics. In: R. Stuewer (ed.) *Minnesota studies in the philosophy of science. Vol. 5. Historical and philosophical perspectives of science* (pp. 311 – 353). Minneapolis: University of Minnesota Press.

Schuman, D. W. & Sales, B. D. (1999). The impact of *Daubert* and its progeny on the admissibility of behavioral and social science evidence. *Psychology, Public Policy, and Law, 5,* 3 – 15.

Shapere, D. (1977). *The structure of scientific theories* (2nd ed.). Chicago: University of Illinois Press

Society for Personality Assessment Board of Trustees (2005). The status of the Rorschach in clinical and forensic practice: An official statement by the Board of Trustees of the Society for Personality Assessment. *Journal of Personality Assessment, 85*, 219 – 237.

Thagard, P. (1978). The best explanation: Criteria for theory choice. *Journal of Philosophy, 75*, 76 – 92.

Thagard, P. (1992). *Conceptual Revolutions.* Princeton, NJ: Princeton University Press.

Vallabhajosula, B. & van Gorp, W. G. (2001). Post-*Daubert* admissibility of scientific evidence on malingering of cognitive deficits. *Journal of the American Academy of Psychiatry and the Law, 29,* 207 – 215.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3,* 231 – 251.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213 – 217.

Weisgram v. Marley Co., 528 U.S. 440 (2000).

Youngstrom, E. A., & Busch, C. P. (2000). Expert testimony in psychology: Ramifications of Supreme Court decision in *Kumho Tire Co., Ltd. v. Carmichael. Ethics & Behavior, 10,* 185 – 193.

Zonana, H. (1995). *Daubert v. Merrell Dow Pharmaceuticals*: A new standard for scientific evidence in the courts? *Bulletin of the American Academy of Psychiatry and the Law, 22,* 309 – 325.